Departamento de Estatística - Universidade de Brasília

Análise de impressões digitais utilizando modelos mistos

Diogo Moreira Chaves Cavalcante

Pedro Henrique Barros Vasconcelos

Brasília

2018

Diogo Moreira Chaves Cavalcante
Pedro Henrique Barros Vasconcelos

Análise de impressões digitais utilizando modelos mistos

Projeto apresentado para obtenção do título de Bacharel em Estatística ao Departamento de Estatística da Universidade de Brasília

Orientador: Prof. Dr. Leandro Tavares Correia

Brasília
2018

Resumo

Este trabalho tem como principal objetivo a utilização de modelos lineares mistos e suas extensões acerca de um banco de dados de impressões digitais com diferentes tratamentos (MALDI-TOF, SALDI-MS, Sem aplicação de micro-líquido LDI) a fim de verificar características semelhantes dos indivíduos, testar a sensibilidade dos métodos utilizados a respeito da capacidade de identificar moléculas endógenas e exógenas e testar o novo tratamento SALDI na análise química de modo que seja possível de fornecer informações relevantes para estudos na área da criminalística. Para a seleção dos modelos foi usado o critério AIC, comparando a significância dos efeitos fixos, quantidades de íons diversas, além de mudanças nas construções da variáveis respostas e nas categorias das variáveis explicativas que foram de melhor proveito na modelagem. O trabalho foi embasado em um conjunto de dados provenientes de uma técnica química chamada espectrometria de massas que foi cedida por um doutorando da área biologia da UnB - Universidade de Brasília. Utilizou-se os pacotes lme4 e nlme no *software R* livre, devido a melhor familiaridade deste.

Palavras-chave: modelos lineares mistos, impressões digitais, espectrometria de massas, *Software R*.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

O termo Criminalística foi primeiramente lançado por Hans Gross para representar o "Sistema de métodos científicos utilizados pela polícia e pelas investigações policiais"(Codeço, 1991). Nas investigações criminais são utilizadas técnicas e metodologias que são baseadas no princípio da troca de Locard e no princípio da individualidade. Este último afirma que mesmo os objetos sendo parecidos eles não são absolutamente idênticos. Já o princípio da troca de Locard afirma que quando dois itens entram em contato eles deixam uma marca um no outro, e que devido ao princípio da individualidade essa marca é tão individual que a partir dela é possível identificar os dois itens.

Um exemplo de marca característica deixada em um contato de objetos é chamado de impressão digital. Essa última é um assinalamento deixado pelo atrito de um dedo humano em um objeto, devido às secreções naturais dos dedos. Em analises químicas mais especificas transcorre um estudo mais complexo das substancias expelidas pela pele, podendo ser estas substâncias endógenas ou exógenas. Quando se pretende realizar perícias ou estudos, essas substâncias são relevantes.

A espectrometria de massas é uma técnica química já difundida em diversas áreas de estudo. A técnica tradicional é muito utilizada para moléculas de tamanho pequeno e médio. Porém, técnicas inovadoras de ionizações utilizando laser permite análises de macromoléculas. Brady et al. (2009), por exemplo, utilizaram dessorção[1] combinada com eletrospray[2] para estudar o espectro de massas de macromoléculas biológicas neutras vindas diretamente do sólido. Nemes e Vertes, 2007, combinaram o laser do infravermelho com a ionização eletrospray para investigar mudanças bioquímicas em organismos in vivo.

Apesar da variedade de técnicas para analisar macro biomoléculas, a grande maioria vem sendo feita através da técnica MALDI-TOF, do inglês Matrix Assisted Lazer Desorption Ionization (MALDI), na qual a amostra é ionizada através de uma ionização por dessorção a laser assistida por matriz, em outras palavras essa técnica seria uma utilização de laser para a liberação das substâncias que se encontram na superfície, e então, os íons são detectados em um analisador do tipo tempo de vôo, do inglês time of flight (TOF). O amplo uso dessa técnica se deve ao fato de que poucas técnicas de diagnósticos microbiológicos tiveram impacto tão rápido em identificação de microrganismos, (Wieser et al., 2012).

Neste experimento, para a obtenção dos dados, foi aplicada a convencional técnica MALDI-TOF e outras duas técnicas, sendo uma sem a utilização de nenhum micro-líquido, e a outra correspondendo a uma nova técnica implementada utilizando o nanomaterial de

[1] utilização de laser para liberar uma substância que se encontra em uma superfície

[2] uma técnica usada para produzir íons em fase gasosa para a análise por Espectrometria de Massas

sílica, que será testada quanto a sua eficiência. Ao se utilizar o espectrômetro, vários íons são obtidos para cada leitura química de impressão digital, ou seja, há mais de uma medida por indivíduo. A estrutura dos dados que essa leitura química gera se torna compatível com a análise via modelos mistos.

Um modelo de efeitos mistos pode ser visto como uma extensão do modelo de regressão clássico correspondente para dados transversais, (Wu, 2009). O modelo de regressão clássico considera que as observações são independentes, porém, nem sempre essa suposição é razoável. O modelo misto leva em consideração a estrutura do conjunto de dados com medidas correlacionadas para os indivíduos, e é com esse tipo de dados que este estudo se baseia.

Em estudos da área criminalística, é escasso a utilização de métodos estatísticos para a análise dos dados, portanto tem-se a devida importância de elaborar esse trabalho prático, visto que uma análise estatística poderia ajudar a encontrar possíveis componentes químicos presentes nas impressões digitais capazes de estarem associados a características fisiológicas, uso de cosméticos ou drogas, entre outros, a fim de ajudar em uma investigação.

Neste trabalho, modelamos e analisamos os dados relativos às 20 coletas de impressões digitais pelos três tratamentos utilizando modelos mistos. A análise será feita de duas formas, uma delas é considerando as intensidades dos íons e a outra é apenas considerando a presença e ausência dos íons. Serão cruzadas informações referentes aos íons das impressões digitais dos indivíduos com características coletadas através de um questionário, além disso, será testado a sensibilidade dos métodos utilizados a respeito da capacidade de identificar moléculas endógenas e exógenas e comparar as técnicas utilizadas, testando o novo tratamento de análise química.

1.1 OBJETIVOS

1.1.1 Objetivo Geral

Analisar e modelar, por meio de modelos mistos, dados de impressões digitais obtidos por meio de experimentos químicos utilizando três tipos de tratamento, verificando assim possíveis características comuns aos indivíduos do estudo.

1.1.2 Objetivos Específicos

- Estudo e revisão bibliográfica a respeito de modelos mistos.
- Estudo de programação do *Software R* para análise dos dados.
- Construir um modelo misto para os dados.
- Verificar possíveis diferenças entre as 3 formas de tratamento das impressões digitais.
- Identificar semelhanças entre indivíduos em relação a alguns aspectos obtidos nos questionários.

1.2 METODOLOGIA

Para a realização da coleta dos dados, foram consideradas algumas informações relevantes ao estudo como sexo, idade, ingestão de cafeína, uso de cosméticos e algumas outras variáveis que foram obtidas por meio da aplicação de um questionário (Anexo 1) e da coleta das impressões digitais dos indivíduos, coleta esta que foi minunciosamente recolhida para que não ocorresse qualquer tipo de viés no estudo, como por exemplo a contaminação das impressões com substâncias externas.

O banco de dados é formado por um grupo de 20 indivíduos, 10 homens e 10 mulheres que tiveram suas impressões digitais submetidos a uma análise em um espectrômetro de massas, com o intento de identificar os diferentes íons que estão presentes, a partir de três tipos diferentes de tratamentos utilizados, o tratamento usual MALDI-TOF, sem tratar as impressões e o novo tratamento de sílica, que será testado quanto a sua eficiência. Cabe destacar que a precisão metodológica estatística está diretamente ligada ao tamanho da amostra, portanto quanto maior a amostra maior será a precisão metodológica e menor o erro amostral, tendo em vista esse ponto, a análise com base nesse grupo de 20 indivíduos é feita, mas o desejável seria que a amostra fosse maior, porém o custo envolvido no experimento é alto.

As abordagens estatísticas que serão utilizadas neste trabalho são a de modelos mistos e a de modelos logísticos mistos, modelos estes que se enquadram na estrutura dos dados. Os dados são estruturados como medidas repetidas para cada pessoa, pelo fato de existirem mais de uma observação para cada impressão digital.

A leitura das impressões digitais foi estruturada de duas formas, sendo a primeira forma considerando a intensidade dos íons quando eles estão presentes na digital e zero quando os íons não aparecem. A segunda forma de estrutura é dada considerando apenas a presença ou ausência dos íons, sendo 1 quando o íon está presente e 0 quando não está.

Para exemplificação de como está especificada a parte relativa às impressões digitais, tem-se um recorte de 5 indivíduos para 6 íons selecionados aleatoriamente no tratamento SALDI.

Tabela 1 – Representação da base de dados considerando a intensidade

Íon	Indivíduo				
	1	2	3	4	5
409	316	116	0	1096	550
125	0	2115	2978	661	0
268	160	332	0	1461	277
111	0	419	3937	1829	656
277	148	179	0	0	300
305	186	146	0	0	0

Tabela 2 – Representação da base de dados considerando a presença ou ausência dos íons

Íon	Indivíduo				
	1	2	3	4	5
409	1	1	0	1	1
125	0	1	1	1	0
268	1	1	0	1	1
111	0	1	1	1	1
277	1	1	0	0	1
305	1	1	0	0	0

Para o primeiro caso, considerando a intensidade dos íons como variável resposta, será utilizado o modelo de efeitos mistos usual, já para o segundo caso, considerando apenas a presença ou ausência dos íons, será utilizado o modelo logístico misto.

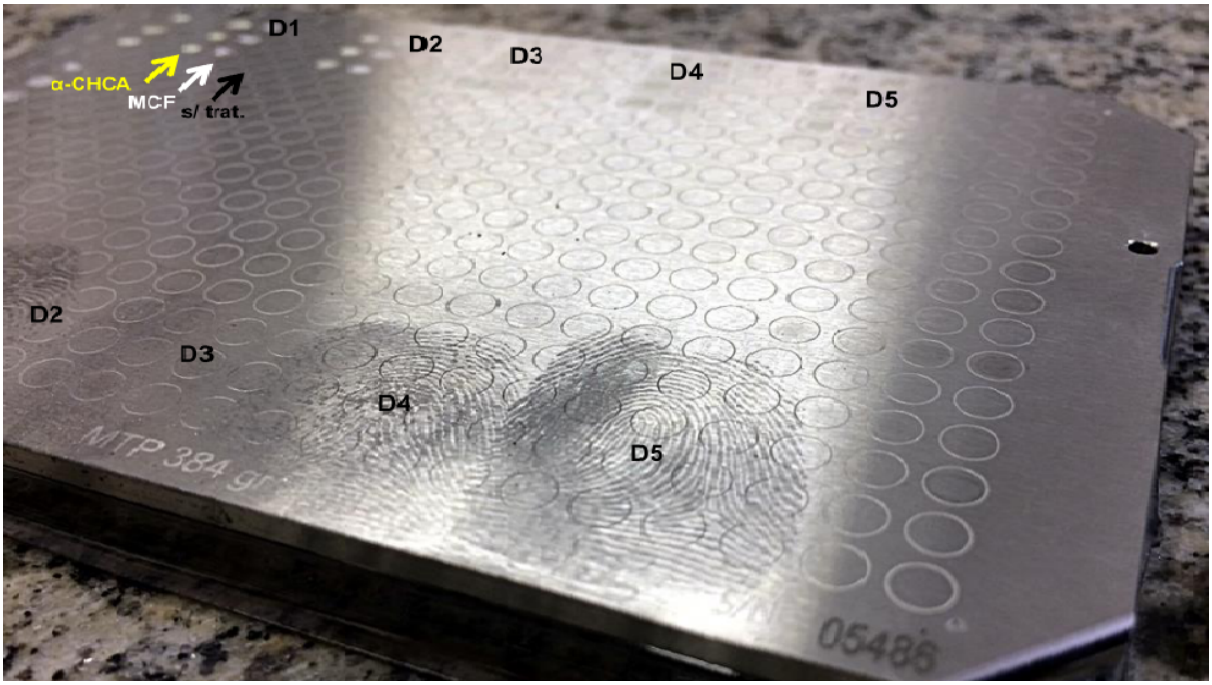
O trabalho será apresentado da seguinte forma, uma revisão bibliográfica a respeito dos temas a serem abordados será exposta nos Capítulos 2 e 3, subsequentemente uma análise descritiva dos dados, a posteriori será feita a modelagem, análise e interpretação dos resultados com o auxílio do software estatístico R.

1.2.1 Espectrometria de Massas

A espectrometria de massa é uma técnica analítica que permite identificar diversos tipos de compostos presentes na amostra a ser analisada, a partir da medição da massa, especificação química da molécula e de seu estado de carga. O espectrômetro de massa ioniza compostos químicos com o objetivo de fragmentar as moléculas e determinar sua razão de massa por carga, por meio de um analisador do tipo tempo de voo, para isso, utiliza-se uma mistura da matriz, ou nanoestruturas como o pó magnético, para facilitar a ionização das moléculas.

Para a realização dessa espectrometria de massas, coleta-se as amostras das digitais em uma placa e são aplicados os tratamentos nessa impressões, como na Figura 1 a seguir:

Figura 1 – Preparação da placa com amostras de impressões digitais submetidas a diferentes tratamentos



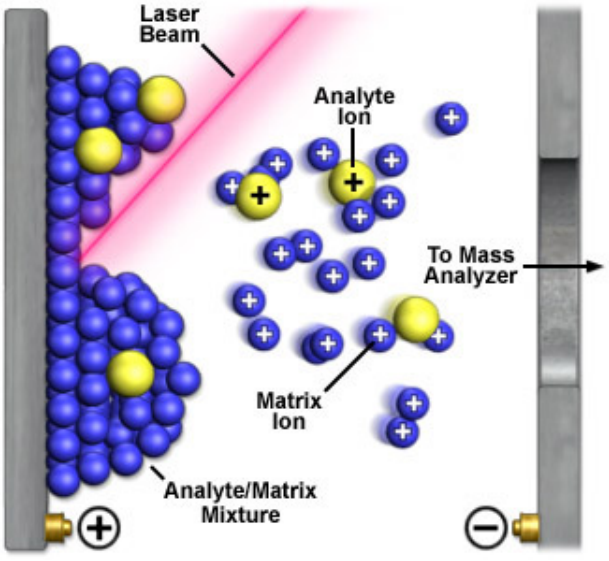
Com as impressões digitais tratadas, o próximo passo será a inserção dessa placa no espectrômetro, gerando assim a leitura química dessas digitais. O espectrômetro é representado na Figura 2 a seguir, assim como uma ilustração de como é dada a ionização das moléculas presentes nessas digitais.

Figura 2 – Espectrômetro de massas

Fonte:http://www.speciation.net/Database/Instruments/Bruker-Daltonics/autoflex-speed-MALDITOFTOF-;i3157

O processo que ocorre na placa com as moléculas presentes se dá conforme a Figura 3. De forma simplificada, as moléculas são bombardeadas com um laser para suas respectivas ionizações, desse modo as moléculas entrarão em estado gasoso e então será possível calcular suas respectivas massas, cargas e a abundância com que os íons estão presentes na impressão.

Figura 3 – Ilustração química da ionização



Fonte:https://prescottbiochem09.wikispaces.com/How+does+MALDI-TOF+work%3F?responseToken=09bf2c9d2f1e49d69f377f47f65e7e2d4

Nessa ilustração, os íons de coloração azul seriam referentes as substâncias presentes nos tratamentos, os íons de coloração amarela seriam as substâncias efetivamente da impressão digital. No caso em que não é aplicado o micro líquido, não se tem estes íons de coloração azul.

2 Modelo de Efeitos Mistos

Os modelos de efeitos mistos são usados principalmente para descrever as relações entre uma variável resposta e algumas covariáveis em dados agrupados de acordo com um ou mais fatores de classificação. Exemplos de tais dados agrupados incluem dados longitudinais, dados multiníveis, projetos de blocos e dados de medidas repetidas, que é o caso dos dados deste trabalho. Ao associar efeitos aleatórios comuns a observações que compartilham o mesmo nível de classificação, os modelos de efeitos mistos representam uma forma flexível de estrutura de correlação induzida pelo agrupamento dos dados.

Os efeitos fixos são os efeitos dos fatores cujos níveis são conhecidos e invariantes no estudo. Não importando a variação do tempo, local, indivíduo e outros, os níveis das variáveis sempre serão os mesmos. Esses fatores estariam representando o comportamento médio dos níveis dos fatores.

Os efeitos aleatórios são os efeitos cujos níveis no estudo constituem uma amostra dos possíveis níveis. São os efeitos específicos de cada indivíduo/cluster em relação à média da população.

Um modelo linear que apresenta somente fatores de efeitos fixos, além do erro experimental, que é aleatório, é denominado modelo de efeito fixo. Os modelos que apresentam apenas fatores aleatórios, exceto a constante β_0, que é fixa, é denominado modelo de efeito aleatório. Os modelos de efeitos mistos são aqueles que apresentam tanto fatores de efeitos fixos como fatores de efeitos aleatórios, além do erro experimental e da constante β_0. Segundo Kuehl (2001), o modelo e a análise para efeitos mistos são compostos de duas partes porque existem dois tipos de inferências. Inferências para o fator de efeitos aleatórios são aplicadas à variação em uma população de efeitos, enquanto inferências para os fatores de efeito fixo são restritos aos níveis específicos usados no experimento.

Para modelos de efeitos mistos, apresentamos efeitos aleatórios para cada indivíduo ou cluster para incorporar a correlação entre as medidas repetidas no indivíduo ou cluster (Wu, 2009).

A técnica do modelo misto é uma criança do casamento das abordagens frequentistas e bayesianas, (Demidenko, 2013). Assim como na abordagem bayesiana, um modelo misto é especificado de forma hierárquica, considerando que os parâmetros são aleatórios mas os hiperparâmetros são estimados a partir dos dados, do mesmo modo que é feito na abordagem frequentista.

Um modelo misto que considera tanto o efeito aleatório quanto o efeito fixo para análise

do experimento é dado em Demidenko (2013) da seguinte forma:

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i \tag{2.1}$$

$$\epsilon_i \sim Normal(0, \sigma_e^2 = \sigma^2 I) \quad b_i \sim Normal(0, \sigma_b^2 = \sigma^2 D)$$

Onde Y_i é o vetor $n_i \times 1$ e que são as respostas repetidas referentes a cada indivíduo i, X_i é a matriz $n_i \times m$ dos dados das variáveis dos efeitos fixos (sendo m o número de efeitos fixos p, adicionado uma unidade, $m = p + 1$), $\beta = (\beta_0, \beta_1, ..., \beta_p)'$ é o vetor $m \times 1$ dos parâmetros da população, que são denominados os efeitos fixos, b_i é o vetor $k \times 1$ dos efeitos aleatórios inerentes a cada cluster i com média igual a 0 e matriz de covariância $\sigma^2 D$, e_i sendo o erro aleatório associado ao modelo com média 0 e matriz de covariância igual a $\sigma^2 I$ onde $I = I_{n_i}$ é uma matriz identidade $n_i \times n_i$ e Z_i a matriz $n_i \times k$ de delineamento de efeitos aleatórios. O fato das variâncias, do erro e do efeito aleatório, apresentarem um σ^2 comum se dá pela notação matricial que será apresentada em (2.6), onde a matriz V faz a união das componentes de variância, e assim as variâncias podem ser expressas como uma variância comum juntamente com suas respectivas componentes.

Um modelo linear misto é considerado balanceado se o tamanho de cada cluster for constante ($n_i = n$) e se a matriz de efeitos aleatórios for a mesma para todos os grupos ($Z_i = Z$). Pela forma como é construída a base de dados, trabalharemos com este tipo de estrutura de dados balanceados.

Em uma forma matricial a equação (2.1) é reescrita na Seção 2.2 em Demidenko (2013) como:

$$Y = X\beta + Zb + \epsilon \tag{2.2}$$

onde

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}, Z = \begin{bmatrix} Z_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & Z_N \end{bmatrix}, b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix} \tag{2.3}$$

e em uma notação simplificada:

$$Y = X\beta + \eta \tag{2.4}$$

em que $E(\eta) = 0$, e η é dado por:

$$\eta = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_N \end{bmatrix} = \begin{bmatrix} \epsilon_1 + Z_1 b_1 \\ \epsilon_2 + Z_2 b_2 \\ \vdots \\ \epsilon_N + Z_N b_N \end{bmatrix} \tag{2.5}$$

A matriz de covariância de η de bloco diagonal $N_T \times N_T$ é expressa em Demidenko (2013) da seguinte forma:

$$V = \sigma^2 \begin{bmatrix} I_{n_1} + Z_1 D Z_1' & 0 & \dots & 0 \\ 0 & I_{n_2} + Z_2 D Z_2' & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & I_{n_N} + Z_N D Z_N' \end{bmatrix} \tag{2.6}$$

por conclusão, o modelo misto (2.1) será descrito de forma genérica como

$$Y_i \sim N(X_i\beta, \sigma^2(I + Z_i D Z_i')). \tag{2.7}$$

A função log-verossimilhança no modelo linear misto é mais elementar ao utilizar a matriz de covariância escalonada dos efeitos aleatórios.

$$D = \frac{1}{\sigma^2} D_* = \frac{1}{\sigma^2} cov(b_i). \tag{2.8}$$

Dispor dessa matriz D nessa parametrização é optado por σ^2 ser melhor analisado, já que se mostra expresso em fórmula fechada, considerando que os outros parâmetros estejam fixos. A matriz de efeitos aleatórios D_* não está escalonada, enquanto a matriz D está.

Para estimar os parâmetros do modelo, duas formas são viáveis: o estimador de máxima verossimilhança (EMV) e o estimador de máxima verossimilhança restrita (EMVR). Esses estimadores serão discutidos nas Seções 2.3 e 2.4.

2.1 Modelos com intercepto aleatório

Os modelos de interceptos aleatórios são os modelos mistos onde se considera que os indivíduos apresentam um padrão de inclinação linear igual para todos estes indivíduos. Para esses modelos, tem-se que os indivíduos não estão correlacionados, há somente a

correlação intra-indivíduo que é a mesma para todas as observações. Dessa forma, a matriz Z é expressa por uma matriz identidade:

$$Z = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{2.9}$$

No exemplo a seguir, é considerado o modelo misto simples (2.10) para ilustrar a correlação introduzida pelo efeito aleatório no modelo, (Wu 2009; Capítulo 2, página 48).

$$y_{ij} = \beta + b_i + e_{ij}, \qquad i = 1, 2, ..., n;\ j = 1, 2, \ldots, m, \tag{2.10}$$

$$b_i \sim Normal(0, \sigma_b^2) \quad \epsilon_{ij} \text{ i.i.d. } \sim Normal(0, \sigma_e^2)$$

Em que o b_i é o efeito aleatório e ϵ_{ij} é o erro aleatório. O efeito b_i adiciona ao modelo a correlação r_i entre as medidas repetidas do indivíduo i. Essa correlação é definida em Wu (2009) da seguinte forma:

$$r_i = Corr(y_{ij}, y_{ik}) = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2}, \quad j \neq k, \quad j, k = 1, 2, ..., m \tag{2.11}$$

Essa correlação permite uma verificação do efeito aleatório. Caso a correlação r_i seja nula, o efeito aleatório b_i seria igual a 0, no entanto, quanto maior a correlação maior seria o efeito aleatório b_i, portanto o modelo misto seria adequado.

2.2 Log-verossimilhança

A estimação comumente usada dos parâmetros para os modelos mistos e para os modelos lineares mistos logísticos se dá basicamente de duas derivações a partir da função log-verossimilhança, são elas: o método de máxima verossimilhança e o método de máxima verossimilhança restrita, o qual serão apresentados a seguir.

Os parâmetros a serem estimados na modelagem de efeitos mistos são β, σ^2 e D, portanto o vetor de parâmetros desconhecidos será $\theta = (\beta', \sigma^2, D)$. A função de log-verossimilhança não dependerá da contante $C = -(N_T/2)\ln(2\pi)$ pois essa constante não afetará na maximização de θ e será omitida para simplificação das equações. A componente N_T representa o número total de observações, sendo o somatório da quantidade de observações de cada cluster, do total de N clusters ($N_T = \sum_{i=1}^{N} n_i$).

Consequentemente a função de log-verossimilhança para o modelo linear misto é descrita em Demidenko (2013) por:

$$l(\theta) = -\frac{1}{2}\{N_T \ln \sigma^2 + \sum_{i=1}^{N}[\ln |I + Z_i D Z_i'| + \sigma^{-2}(y_i - X_i\beta)'(I + Z_i D Z_i')^{-1}(y_i - X_i\beta)]\}, \tag{2.12}$$

ou em uma notação simplificada

$$l(\theta) = -\frac{1}{2}\{N_T \ln \sigma^2 + \sum_{i=1}^{N}\left[\ln |V_i| + \sigma^{-2} e_i' V_i^{-1} e_i\right]\}, \tag{2.13}$$

onde V_i representa a razão do i-ésimo termo da diagonal da matriz V por σ^2

$$V_i = I + Z_i D Z_i', \tag{2.14}$$

e o vetor de resíduos $n_i \times 1$ para o i-ésimo cluster é

$$e_i = y_i - X_i\beta. \tag{2.15}$$

2.3 Estimador de Máxima Verossimilhança

A metodologia de máxima verossimilhança constitui-se em estimar parâmetros de um modelo pela maximização da função de log-verossimilhança, tornando, dentre todas as estimativas possíveis, as estimativas com maior probabilidade de ter gerado o conjunto de valores da amostra.

A maximização da função log-verossimilhança (2.12) em relação a σ^2 é demonstrada na Seção 2.2 em Demidenko (2013). A função é máxima quando:

$$\hat{\sigma}^2 = \frac{1}{N_T}\sum_{i=1}^{N}(y_i - X_i\beta)'(I + Z_i D Z_i')^{-1}(y_i - X_i\beta) \tag{2.16}$$

O estimador para o vetor de parâmetros β é dado por

$$\hat{\beta} = \left[\sum_{i=1}^{N} X_i'(I + Z_i D Z_i')^{-1} X_i\right]^{-1}\left[\sum_{i=1}^{N} X_i'(I + Z_i D Z_i')^{-1} y_i\right] \tag{2.17}$$

e sua demonstração está especificada em Demidenko (2013); Capítulo 2, onde é utilizado redução de dimensão para encontrar a estimativa que maximize a função de log-verossimilhança.

As expressões acima são formas recorrentes, diferente de como seria em um modelo de regressão clássico. A resolução das fórmulas se dá de forma iterativa, uma vez que o valor estimado do vetor de β em (2.17) depende do valor estimado de σ^2 na equação (2.16), que por sua vez depende do valor estimado de β, e após repetido esse processo convergir para

um ponto de máxima. Para suas respectivas aproximações são necessários a realização de passos, onde a cada iteração se gera um valor para o estimador de máxima verossimilhança do parâmetro, e que se espera a convergência para o valor real do parâmetro estimado.

2.4 Estimador de Máxima Verossimilhança Restrito

Esse método é uma extensão do estimador de máxima verossimilhança que tem por objetivo, além de estimar os efeitos fixos, corrigir o viés dos componentes da variância do estimador de máxima verossimilhança. A estimativa obtida pelo método de máxima verossimilhança para os componentes de variância em amostras finitas apesar de ser consistente é viesada e subestima as estimativas dos parâmetros.

(Patterson e Thompson, 1971) propuseram uma mudança na log-verossimilhança para contemplar a perda dos graus de liberdade devida aos efeitos fixos e assim retirar o viés do estimador, assim esse novo estimador foi denominado como o de máxima verossimilhança restrito. A forma da nova verossimilhança restrita é descrita como:

$$l_{MR}(\theta) = \left[\sum_{i=1}^{N} X_i'(I + Z_i D Z_i')^{-1} X_i\right]^{-\frac{1}{2}} l(\theta) \tag{2.18}$$

onde $l(\theta)$ é a log-verossimilhança definida em (2.12). A nova estimativa para a componente de variância com as devidas alterações feitas é representado da seguinte forma:

$$\hat{\sigma}_R^2 = \frac{1}{N_T - m} \sum_{i=1}^{N} (y_i - X_i\beta)'(I + Z_i D Z_i')^{-1}(y_i - X_i\beta) \tag{2.19}$$

O método da máxima verossimilhança restrita também é iterativo e exige os mesmos pressupostos do método de máxima verossimilhança e sua estimação seguirá conforme os métodos iterativos de otimização.

A alteração somente se encontra na estimação de σ^2, o vetor de efeitos fixos β permanece da mesma forma e sua estimação não é alterada.

O modo como a matriz de covariância é construído interfere diretamente nos resultados obtidos ao se modelar os dados. A estrutura apresentada neste trabalho é bastante utilizada.

2.4.1 Critério AIC

Como norma de decisão quando se trata de qual estrutura de covariância é mais adequada para a análise, um critério de ajustes dos modelos bastante utilizado e com resultados confiáveis é o critério de Informação de Akaike (AIC). Trata-se de valores do logarítmo da função de Máxima Verossimilhança penalizados pelo número de parâmetros.

Quanto menor o AIC em valor absoluto, este é preferível para a escolha do melhor modelo.

$$AIC = -2l(\theta) + 2[(p+1)+1], \tag{2.20}$$

onde p é o número de parâmetros do modelo e $l(\theta)$ é o máximo do logaritmo da verossimilhança que será presentado no capítulo a seguir.

2.5 Algoritmos de Maximização

Em muitos casos na estimação dos parâmetros do modelo, somente uma fórmula recursiva é possível de ser expressa para os estimadores, um exemplo disso seria a equação dada em (2.16). Face ao exposto, algoritmos de maximização são necessários para uma aproximação dessas estimativas e assim se obter um valor ótimo para a função de log-verossimilhança, alcançando assim estimadores para o modelo (2.1).

Um algoritmo de maximização é um algoritmo iterativo usado para encontrar soluções aproximadas para problemas de otimização. Os três algoritmos mais comumente usados na estatística para a maximização da função de log-verossimilhança são o de Newton-Raphson (NR), Fisher scoring (FS) e Expectation-Maximization (EM), onde existem vantagens e desvantagens comparativamente a cada algoritmo, que serão explicadas posteriormente. Estes três algoritmos são similares quanto aos fundamentos principais, eles apresentam uma forma comum que podem ser expressa como

$$t_{s+1} = t_s + \lambda_s \delta_s, \quad s = 0, 1, 2, ... \tag{2.21}$$

onde t_s representa a estimativa da função, λ_s indica o tamanho do passo e δ_s demonstra a direção para onde a função irá, todos na s-ésima iteração. A partir desses valores uma nova aproximação pode ser feita para o máximo da log-verossimilhança. A construção da variável δ_s é o diferencial de cada algoritmo, podendo ser escrito em uma fórmula generalizada como

$$\delta_s = H_s^{-1} \frac{\partial l}{\partial t}\bigg|_{t=t_s} \tag{2.22}$$

Mais especificamente, cada algoritmo difere quanto a matriz H_s. A escolha de como será a forma dessa matriz H_s define cada algoritmo de maximização.

2.5.1 Expectation-Maximization (EM)

Na computação estatística, o algoritmo de expectativa máxima (EM) é um algoritmo iterativo projetado para encontrar a estimativa de máxima verossimilhança dos parâmetros, normalmente é utilizado quando temos dados não observáveis.

Segundo (Casella e Berger, 2010), o EM é um algoritmo que seguramente converge para o estimador de máxima verossimilhança e tem como base a ideia de substituir uma dificil maximização da verossimilhança por uma sequência de maximizações mais fáceis, cujo limite é a resposta para o problema original.

Para o modelo de efeito misto, as iterações EM são baseadas em relação ao efeito aleatório, já que o $b_i, i = 1, \ldots, M$, são dados não observados. Usamos o vetor de parâmetros de variancia-covariância vigente na iteração w, $\theta(w)$, para avaliar a distribuição de $b|Y = y$ e derivar a esperança da probabilidade de log-verossimilhança para um novo valor de θ, dada esta distribuição condicional. Como se apresentou uma esperança, esse passo é chamado de passo E. O passo M consiste em maximizar essa esperança em relação a θ para produzir $\theta(w + 1)$. Cada iteração do algoritmo EM resulta em um aumento na função log-verossimilhança, mesmo que seja, possivelmente, um pequeno aumento, a função estará mais próxima do máximo, gerando uma melhor estimação dos parâmetros para o modelo em questão.

2.5.2 Newton-Raphson (NR)

No algoritmo de Newton-Raphson, H é determinada pela matriz hessiana de segundas derivadas. O passo λ_s segue fixado da seguinte forma: λ_s é igual a 1 inicialmente, se $l_{s+1} > l_s$ considerar $\lambda_s = 1$, caso contrário considerar $\lambda = 1/2, 1/2^2, \ldots$ até $l_{s+1} > l_s$ e então prosseguir para a próxima iteração. λ_s pode não ser positivo em dois casos, quando o gradiente é próximo de zero, nessa situação, em via de regra, o máximo local é encontrado e é determinada a parada do algoritmo. A outra possibilidade de λ_s não ser positivo é quando a matriz hessiana não é definida positivamente e nesse caso o algoritmo falha, algo que ocorre principalmente quando o ponto de partida é longe do ponto de máximo. Essa é a grande desvantagem desse algoritmo.

Esse algoritmo converge em velocidade quadrática para o máximo da função se o ponto de partida escolhido for relativamente próximo do máximo, porém pode falhar se o ponto de partida estiver longe do máximo.

As condições para o sistema de equações alvo são apenas que elas sejam diferenciáveis na região e o sinal relativo à segunda derivada.

2.5.3 Fisher scoring (FS)

O algoritmo de Fisher scoring possui a velocidade de convergência muito próxima do Newton-Raphson, mas é mais confiável por não ser afetada por *outliers* e pelo fato de a matriz H ser a matriz de informação de Fisher, determinada pelo valor esperado do inverso da matriz hessiana, que sempre é definida positiva dado que o modelo esteja bem

definido. O fato de a matriz ser definida positivamente implica que esse algoritmo possui maior robustez em relação a escolha do ponto de partida.

2.6 Inferência para os Parâmetros de Efeitos Fixos

A distribuição dos parâmetros dos efeitos fixos são baseados na matriz de variância e covariância, ao qual seu estimador é dado de forma assintótica. A matriz de covariância, do vetor de $\hat{\beta}$, é descrito no Capítulo 3 em Demidenko (2013), e assume a forma aproximada de:

$$\widehat{cov(\hat{\beta})} = \hat{\sigma}^2 \left(\sum_{i=1}^{N} X_i'(I + Z_i \hat{D} Z_i')^{-1} X_i \right)^{-1} \tag{2.23}$$

Também é descrito na Seção 3.4 em Demidenko (2013), a distribuição de $\hat{\beta}_j / s_j$, que é assintoticamente t-student com $N_t - m$ graus de liberdade onde s_j é a raiz quadrada do j-ésimo elemento da diagonal da matriz de covariância de beta, o qual representa o erro padrão do $\hat{\beta}_j$ a ser testado. Partindo dessa distribuição é possível testar a significância do parâmetros de efeitos fixos.

Para o teste da significância dos efeitos fixos, ou seja, testar se esses efeitos são iguais a zero, tem-se as seguintes hipóteses:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_A : \beta_j \neq 0 \end{cases} \tag{2.24}$$

A função *lme* do *Software R* considera as estimativas por máxima verossimilhança restrita e desse modo a estatística do teste é t-Student. A correspondente estatística do teste é:

$$T_0 = \frac{\hat{\beta}_j}{EP(\hat{\beta}_j)} \sim t_{(N_t - m)} \tag{2.25}$$

com distribuição t-Student com $N_t - m$ graus de liberdade.

3 Modelos Lineares Generalizados

O universo de análise de regressão é muito maior do que o modelo de regressão clássico, existem várias outras extensões do modelo de regressão linear, onde uma classe de extensão é conhecida como Modelos Lineares Generalizados (MLG's).

Os Modelos Lineares Generalizados (MLG's) conseguem tratar as observações da variável reposta quando são variáveis aleatórias independentes com distribuição de probabilidade pertencente a família exponencial. Dos quais só se tornam possíveis porque existem funções de ligações especificas e respectivas para cada tipo de distribuição da variável resposta.

Nesse contexto de modelagem, em um universo muito mais amplo, em que a distribuição de probabilidade dos dados não está simplesmente restrita a distribuição normal, um deles para se trabalhar com dados na categoria de binários, isto é, onde exitem apenas duas respostas possíveis, "sucesso"e "fracasso", seria considerar a distribuição Binomial. Para o caso de considerar a distribuição Binomial nota-se a necessidade de se lidar com o modelo de regressão logística, que é um caso particular de MLG's que usa a transformação logito como função de ligação.

3.1 Modelo Linear Misto Generalizado (GLMM)

O preceito trazido pelo modelo linear generalizado consegue tratar unicamente estudos que possuem variáveis de efeito fixo. Uma expansão dessa ideia para que seja possível trabalhar com variáveis que não sigam este princípio seria a utilização de modelos que consigam ajustar a dados coletados de forma experimental onde os níveis de determinado fator já foram previamente selecionados a partir de uma população já ordenada por níveis, ou seja, com níveis aleatórios.

A inserção de efeitos aleatórios nos modelos lineares generalizados revela um outro modelo que é conhecido como modelo linear misto generalizado, possibilitando assim, junto com as covariáveis agregadas ao experimento, a construção de uma modelagem na estrutura de correlação entre as observações que fazem parte do mesmo cluster. Portanto, o GLMM estendem muito a aplicabilidade dos modelos lineares mistos.

A função de ligação é um importante componente dos modelos generalizados. Ela indica a relação adotada entre as variáveis explicativas, de modo linear, e a variável resposta, em média.

A seleção de funções de ligação deve estar necessariamente de acordo com o que se propõe para a análise de casos de pesquisa específicos.

3.1.1 Regressão Logística Mista

Sumariamente, segundo Demidenko (2013), o modelo linear misto generalizado é uma extensão do modelo linear generalizado (GLM) adicionado o efeito aleatório. Partindo dessa afirmação, para o caso específico da regressão logística, um modelo de efeitos mistos é construído pela introdução do efeito aleatório no modelo de regressão logístico padrão. O modelo linear misto logístico resultante é:

$$\ln\left[\frac{p_{ij}}{1-p_{ij}}\right] = X_{ij}\beta + Z_i b_i + \epsilon_{ij} \tag{3.1}$$

$$b_i \sim N(0, \sigma_b^2)$$

Onde, p_{ij} é a probabilidade da indivíduo i ter a observação j em questão, β é o vetor de parâmetros fixos do modelo, X_{ij} são os valores da matriz de variáveis explicativas X para a observação j do indivíduo i e Z_i é a matriz de delineamento de efeitos aleatórios. O efeito aleatório continua sendo o b_i e com a suposição de distribuição normal, $b_i \sim Normal(0, \sigma_b^2)$. Além dessas componentes, tem-se o erro ϵ_{ij}, e que nenhuma suposição é feita sobre a distribuição de probabilidade deste erro.

A esperança da variável resposta é dado da seguinte forma:

$$E(y_{ij}) = P(y_{ij} = 1) \tag{3.2}$$

e a variância é:

$$\mathrm{Var}(y_{ij}) = P(y_{ij} = 1)P(y_{ij} = 0), \tag{3.3}$$

Essa probabilidade de "sucesso"ou "fracasso"está relacionada ao vetor de parâmetros por meio da função de ligação, pois considerando que agora a variável resposta terá distribuição binomial. O parâmetro correspondente a probabilidade de sucesso da binomial será o valor esperado das respostas definido por $P(y_{ij} = 1)$.

O modelo de regressão misto logístico pode ser especificado como y_i sendo o vetor $(n_i \times 1)$ da variável dependente para o indivíduo i e x_i é a matriz $(n_i \times m)$ de observações das covariáveis para o indivíduo i, $i = 1, \ldots, N$. É assumido que o número de observações é maior ou igual ao número de covariáveis $(N \geq m)$ e que a matriz X com as N matrizes x_i tem posto completo.

De forma análoga ao que foi visto na seção 2.2, a log-verossimilhança para o caso da regressão mista logística é definida da seguinte forma:

$$l(\beta) = \sum_{i=1}^{n} \left[y_i \ln P(y_i = 1) + (1 - y_i) \ln P(y_i = 0)\right] \tag{3.4}$$

e todos os passos para se chegar nessa forma podem ser vistos em (Demidenko 2013, Capítulo 7).

Uma explicação sobre a maximização da log-verossimilhança para a estimação dos parâmetros por máxima verossimilhança seguirá de maneira similar ao visto na seção 2.3, com a diferença de que a verossimilhança irá mudar porque é assumido uma outra distribuição de probabilidade.

Segundo Demidenko (2013), Seção 7.1.4, o estimador de máxima verossimilhança não terá fórmula fechada. O máximo da função de log-verossimilhança é atingido quando a derivada for igual a zero, ou seja:

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{N} (y_i - P(y_i = 1)) \frac{P(y_i = 1)}{P(y_i = 1)P(y_i = 0)x_i} = 0 \tag{3.5}$$

e como visto anteriormente na Seção 2.4, uma penalização também se faz necessária para retirar o viés do estimador para o σ^2, essa penalização é dada de forma análogo ao que foi feito no modelo de regressão linear misto. Feito essa penalização, tem-se o novo estimador de máxima verossimilhança, agora denominado estimador de máxima verossimilhança restrito.

Considerando o que já foi comentado na Seção 2.4, o estimador de máxima verossimilhança restrito possui fórmulas recursivas, tornando-se necessária a utilização de algoritmos de maximização para se alcançar uma melhor estimativa dos parâmetros. Equivalente ao que foi visto na Seção 2.5, os mesmos algoritmos se aplicam e a estimação dos parâmetros será dada da mesma forma.

O teste de significância para os efeitos fixos é dada de uma forma diferente ao apresentado na Seção 2.6. Na regressão logística mista, utiliza-se o teste de Wald para testar essa significância, e de forma simplificada, as hipóteses do teste é dado da seguinte forma:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_A : \beta_j \neq 0 \end{cases} \tag{3.6}$$

A função *glmer* do *Software R* considera as estimativas por máxima verossimilhança para os efeitos fixos e desse modo a estatística do teste será assintoticamente normal padrão. A correspondente estatística do teste é:

$$W_0 = \frac{\hat{\beta}_j}{\widehat{EP(\hat{\beta}_j)}} \sim Normal(0, 1) \tag{3.7}$$

em que o $\widehat{EP(\hat{\beta})}$ é o valor estimado do erro padrão do estimador.

4 Aplicação

4.1 Análise Descritiva

O banco de dados é dividido em duas partes, sendo a primeira relativa aos questionários respondidos dos 20 indivíduos e a segunda parte referente a análise química obtida pelo espectrômetro de massa.

O questionário aplicado possui 9 perguntas, no entanto em uma pergunta (se possui restrição na dieta) a resposta foi a mesma para todos os indivíduos, portanto essa variável foi desconsiderada na modelagem. As $k = 8$ variáveis que serão estudadas do questionário são: gênero (masculino e feminino), idade, quantidade de cigarros fumados por dia, se faz uso frequente de algum medicamento (cita-lo caso a resposta for sim), se o indivíduo tomou banho pela manhã antes de doar as impressões digitais, se a pessoa utilizou cosméticos nas horas que antecederam o exame (e cita-los caso a resposta for sim), se utiliza produtos de limpeza frequentemente (essa variável será denotada como limpeza) e se ingeriu café na manhã antes de doar suas impressões digitais. O questionário aplicado está em anexo no relatório.

Para fins de modelagem e análise dos dados, a variável uso de cosmético, foi subdividida em duas variáveis, sendo a primeira: utilizou creme ou protetor solar ou perfume ou desodorante ou maquiagem (CPPDM) e a segunda: uso de xampu ou condicionador ou sabonete (XCS). Essa separação se dá de forma intuitiva, pois no primeiro caso são produtos que entram em contato direto com a pele do doador e no segundo caso são produtos de limpeza corporal.

Dos doadores da amostra, 10 pessoas são homens e 10 são mulheres entre 20 e 52 anos, desses 20 indivíduos, apenas 2 fumam, 8 afirmaram fazer uso de algum medicamento, os medicamentos foram: dipirona sódica, anticoncepcional e antidepressivo, 10 tomaram banho nas horas que antecederam a coleta da digital, 14 utilizaram ou xampu ou condicionador ou sabonete, 14 usaram ou creme ou protetor solar ou perfume ou desodorante ou maquiagem, 15 têm contato com o produto de limpeza frequentemente e 10 tomam café. Algumas representações tabulares e gráficas, para uma melhor visualização dessas informações podem ser vistas na Figura 4 e nas Tabelas 3, 4 e 5 a seguir.

Tabela 3 – Número de pessoas por sexo

Sexo	Frequência
Masculino	10
Feminino	10

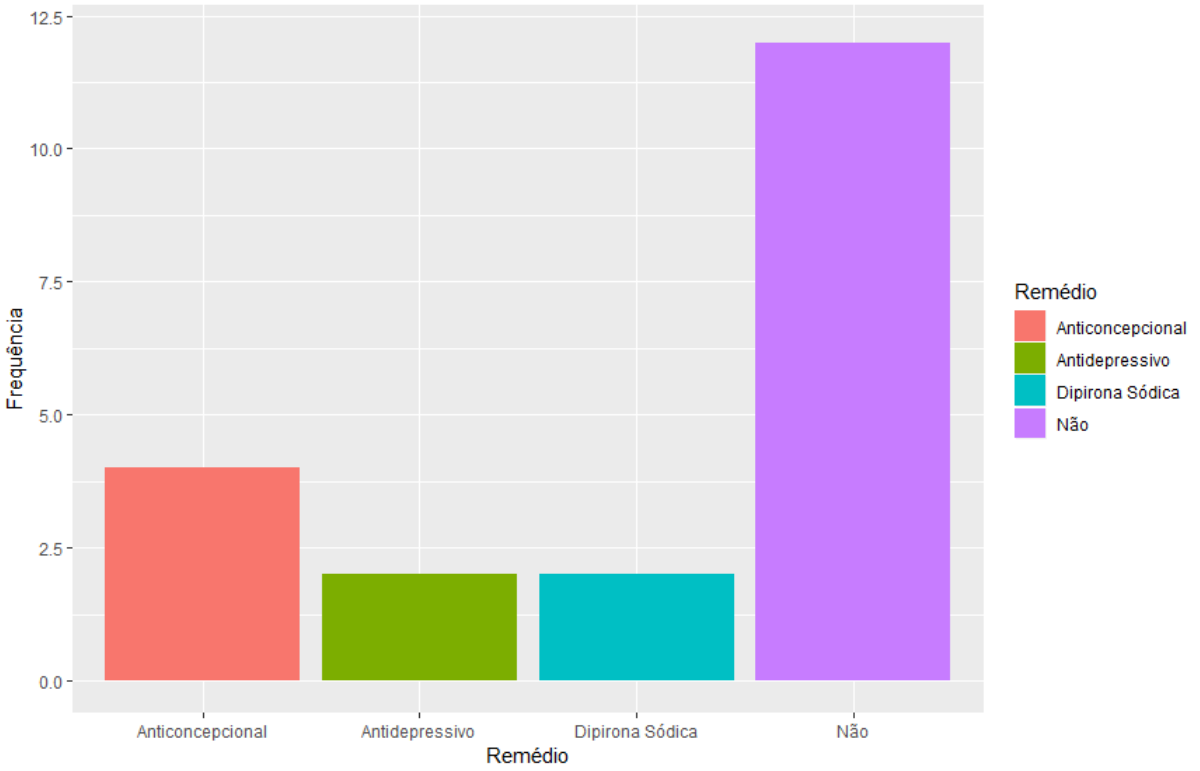
Tabela 4 – Distribuição da faixa etária por sexo

Sexo	Faixa Etária
Masculino	24 |- 52
Feminino	20 |- 32

Tabela 5 – Frequência de respostas positivas para as variáveis do questionário

Variável	Frequência
Fumante	2
Banho	10
Café	10
XCS	14
CPPDM	14
Limpeza	15

Figura 4 – Frequência dos tipos de medicamentos



Na amostra coletada, de todas as dez pessoas que ingeriram café antes de doar suas digitais, apenas uma não têm contato com produto de limpeza habitualmente e apenas uma não tomou banho, enquanto que das pessoas que tomaram café, apenas uma tomou banho. De todos os dez indivíduos que não tomaram banho pela manhã, apenas um toma café e os outros dez indivíduos que tomaram banho, apenas um não tomou café e não tem contato com produto de limpeza frequentemente. Dos dez homens somente um faz uso

frequente de medicamento e de todos indivíduos que fazem uso de medicamento, somente uma não tem contato com produto de limpeza comumente.

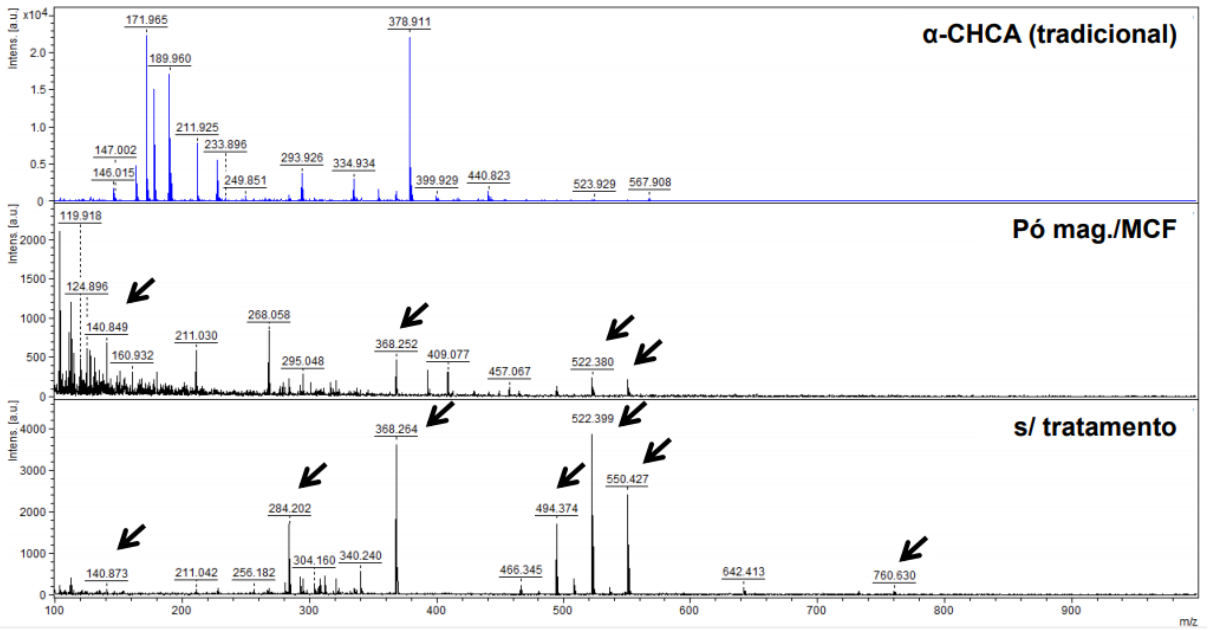
Para coletar as impressões digitais de cada pessoa, se teve bastante cuidado com a limpeza da placa para que não houvesse interferência de fatores externos ao de interesse do estudo.

Após a análise química dessas impressões digitais pela técnica de espectrometria de massas nos três tratamentos, é observado a taxa massa por carga (m/z) e a abundância dos íons presentes na amostra. Cada taxa (m/z) significa uma identidade com algum íon, sendo essa taxa, mesmo em uma escala numérica, um identificador de um determinado íon, por exemplo o íon com a taxa 466 que representa a substância Sufactante, substância essa presente em detergentes e produtos de limpeza em geral, logo essa taxa representa uma variável qualitativa ao passo que a taxa é somente a identificação de cada íon, enquanto a variável intensidade é analisada como uma variável quantitativa.

Em um caso específico da amostra para o tratamento MALDI, temos os íons 249.85 e 250, que apresentam as razões de massa por carga (m/z) muito próximas. Destaca-se o fato de que esses íons podem não estar correlacionados, portanto o fato de considerarmos essa medida como qualitativa se fundamenta.

Na Figura 5 que segue, temos a representação gráfica dos íons, com suas respectivas intensidades, para um doador nos três tratamentos.

Figura 5 – Espectros representativos



Uma característica importante dos dados referentes aos íons dos doadores da amostra é que, além de serem várias observações de íons para cada doador, temos uma quantidade

diferente de íons por pessoa. Para contornar esse problema de dados desbalanceados, consideraremos somente os íons com maior frequência de observações nos indivíduos. Dado a escolha dos íons de interesse na modelagem, nem todas a impressões digitais terão esses íons, nesses casos a resposta de intensidade y_{ij} será igual a 0, desse modo teríamos dados balanceados, ou seja, onde a quantidade de medidas repetidas seria igual para todos os indivíduos.

Ao se considerar que vários íons são exclusivos de certos indivíduos, íons esses que corresponderiam a características específicas dessas pessoas, optou-se por considerar apenas os íons que são comuns a 5 ou mais indivíduos da amostra. Esse ponto de corte seria suficiente para obter informações relevantes e desprezar informações que não contribuem para a análise.

Em relação aos íons comuns a todos os tratamentos, os que se mostraram presentes no MALDI, SALDI e sem aplicação de líquido foram: 178, 268, 284, 295, 312, 494, 522, 550.

4.1.1 α-matriz (MALDI-TOF)

As impressões digitais que foram tratadas com α-matriz obtiveram uma maior quantidade de íons ionizados com intensidade significativa, os íons das tabelas a seguir foram os que apareceram em cinco ou mais impressões digitais e serão consideradas na posterior modelagem estatística.

Tabela 6 – Quantidade de Ionizações para o tratamento MALDI-TOF

Íon	107	121	122	123	208	249.85	261	278	291	309	395
Frequência	5	5	5	5	5	5	5	5	5	5	5

Íon	439	455	506	666	152	194	247	274	333	337	362	413
Frequência	5	5	5	5	6	6	6	6	6	6	6	6

Íon	433	443	471	138	191	268	282	130	207	312	128	295
Frequência	6	6	6	7	7	7	8	9	9	9	10	10

Íon	304	568	139	256	494	147	691	189	192	400	227	441
Frequência	10	10	11	11	11	12	12	13	13	13	14	14

Íon	104	146	401	417	550	250	335	266	294	164	190	234
Frequência	15	15	15	15	15	16	17	18	18	19	19	19

Íon	284	522	172	178	212	228	354	379
Frequência	19	19	20	20	20	20	20	20

Neste tratamento notou-se a presença de uma quantidade maior de íons do que sem a aplicação de líquido ou com a aplicação da sílica, isso ocorre principalmente devido

a molécula do material da α-matriz supostamente facilitar a ionização das moléculas presentes na amostra, e também pelo seu material poder ser ionizado, esse último caso seria um resquício não desejável e que dificultaria na análise dessas impressões digitais.

Os íons presentes em todas as impressões digitais são: 172, 178, 212, 228, 354 e 379, as intensidades desses íons apresentaram as seguintes médias, desvios padrões e coeficientes de variação:

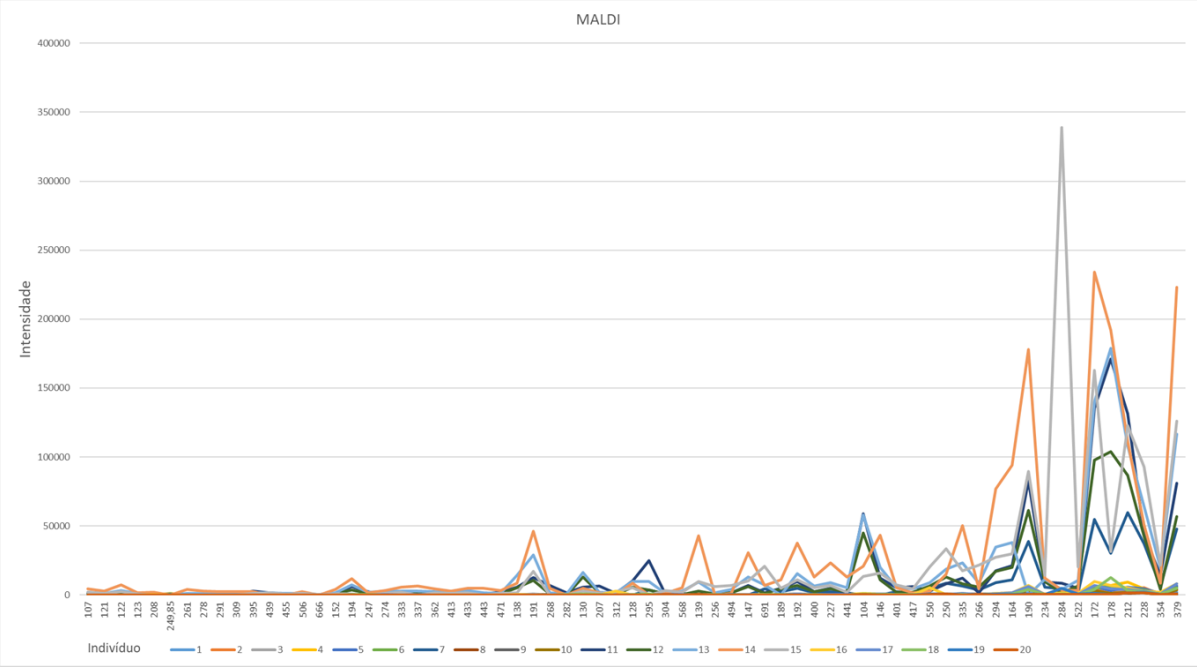
Tabela 7 – Medidas Resumo da amostra para o Tratamento MALDI

íon	172	178	212	228	354	379
Média	44019.10	37743.75	33499.75	18584.95	4159.00	35363.30
Desvio Padrão	70091.01	65900.97	48777.04	26746.12	6628.18	59172.98
C_v	1.5923	1.7460	1.4560	1.4391	1.5937	1.6733

Os íons que estão presentes somente no MALDI e sem aplicação do líquido são: 121, 191, 309, 335, 354, 433.

Algo relevante desse tratamento é que todos os seis indivíduos que usam produto de limpeza frequentemente e não tomaram banho pela manhã, apresentaram uma quantidade maior de ionizações nas impressões digitais comparativamente aos demais indivíduos que tomaram banho ou que não fazem uso frequente de produto de limpeza, além de terem uma intensidade dos íons, em média, cerca de 34 vezes maior.

Figura 6 – Gráfico de perfil para a intensidade dos íons

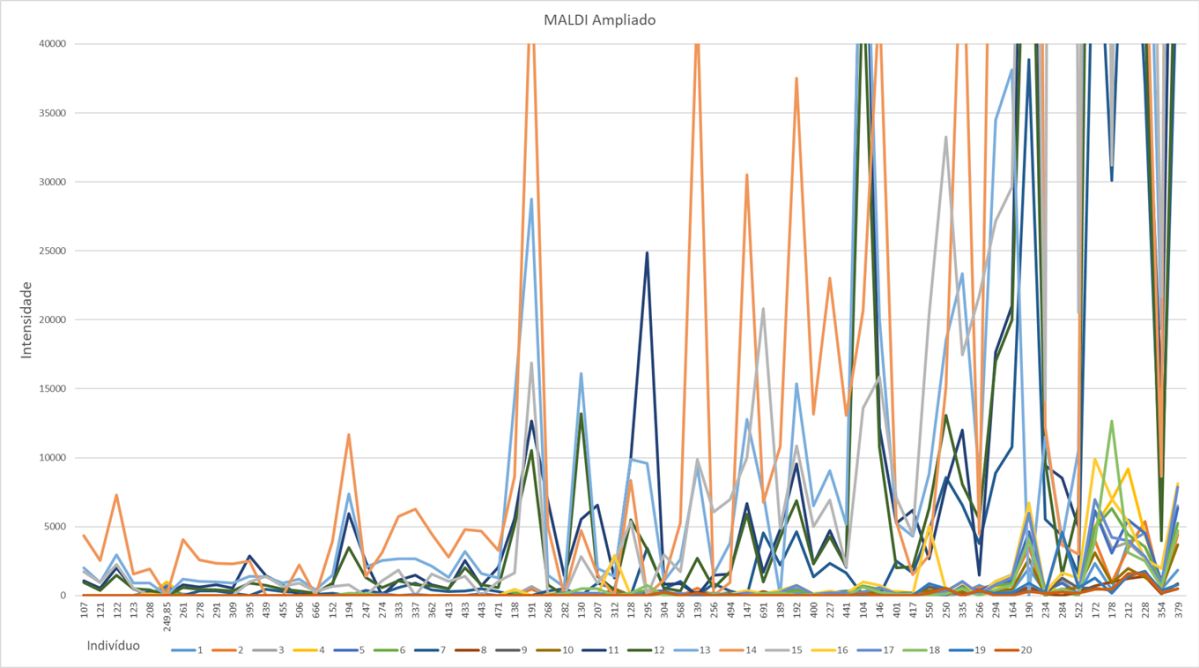


No gráfico da Figura 6, os íons estão ordenados pela frequência em que aparecem nos indivíduos da amostra, sendo os íons que aparecem em apenas 5 indivíduos na esquerda e os íons que aparecem para todos os indivíduos na direita.

Algo que se destaca nesse gráfico de perfil (Figura 6) são os seis indivíduos que apresentam comportamento destacado e similar em ralação a picos de intensidade em alguns íons: 104, 122, 128, 130, 146, 147, 164, 178, 191, 192, 194, 212, 227, 228, 250, 266, 294, 335, 379, 400, 401, 417, 441, 550, 691, esses indivíduos são: 7, 11, 12, 13, 14, 15. Esses seis indivíduos responderam não para a variável banho e sim para a variável limpeza. Essas seis pessoas são as únicas a terem essa combinação de resposta no questionário.

Devido a alta intensidade de alguns íons dessas seis impressões digitais relativamente as outras impressões, a visão das faixas de intensidades mais baixas fica limitada e uma imagem ampliada facilita a análise visual do restante dos indivíduos (vide Figura 7).

Figura 7 – Gráfico de perfil para a intensidade dos íons



Nesse gráfico apresentado (Figura 7), para os demais indivíduos que não se destacam pela alta intensidade dos íons, não é possível visualizar padrões de intensidades correlacionadas com as respostas dos questionários.

4.1.2 Pó magnético/Sílica (SALDI-MS)

O novo método Surface-assisted laser desorption/ionization (SALDI), a base de sílica, apresentou um número maior de ionizações comparado ao método sem aplicação de micro-líquido, isso se dá ao fato de ter uma interação da sílica com as moléculas presentes nas impressões digitais, auxiliando assim suas respectivas ionizações. Ao se comparar a nova técnica com a tradicional MALDI, tem-se que o número de ionizações foi menor, o que pode indicar um menor ruído alusivo a interação das moléculas com o material utilizado para a facilitação da ionização.

A seguir está representada as tabelas com os íons com frequência maior ou igual a cinco nas leituras das impressões digitais:

Tabela 8 – Quantidade de Ionizações para o tratamento SALDI-MS

Íon	138	307	323	429	135	167	508	116	312	368	536
Frequência	5	5	5	5	6	6	6	7	7	7	7

Íon	113.1	112	124	129	178	211	319	120	151	157	409
Frequência	7	8	8	8	8	8	8	9	9	9	9

Íon	125	268	111	277	305	308	393	113	141	279	284
Frequência	10	10	11	11	11	11	12	13	13	14	15

Íon	293	128	321	494	104	115	295	550	522	113
Frequência	15	17	17	17	18	19	19	19	20	20

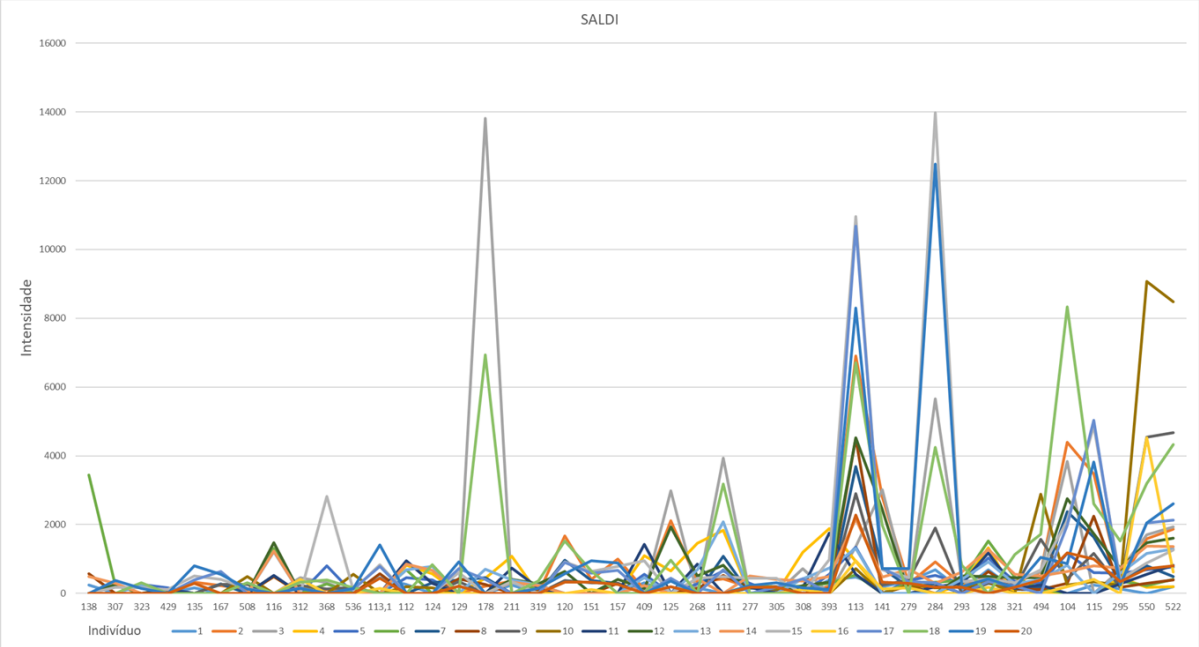
Os únicos íons presentes em todas as impressões digitais para o tratamento a base de sílica foram os íons 522 e 113, as intensidades desses íons apresentaram as seguintes médias, desvios padrões e coeficientes de variação:

Tabela 9 – Medidas Resumo para o Tratamento SALDI

íon	522	113
Média	1810.90	3590.80
Desvio Padrão	2001.63	3396.46
C_v	1.105	0.946

O tratamento SALDI destacou alguns íons também presentes no tratamento MALDI (104, 128, 138).

Figura 8 – Gráfico de perfil para a intensidade dos íons



No caso desse novo material que está sendo testado, conforme a Figura 8, este se mostrou graficamente constante em relação a ter intensidades baixas dos íons, exceto quando se visualiza os íons presentes para quase todas as pessoas e em cinco picos altos específicos, nos íons (104, 111, 113, 178, 284), onde dois indivíduos (3 e 18) têm um comportamento próximo nesses picos. Outros dois indivíduos também apresentam um comportamento similar foram o 15 e o 19, porém com intensidades altas somente nos íons: 113, 115, 284. Cabe ressaltar que não foi identificada semelhança entre estes indivíduos em relação as perguntas do questionário.

4.1.3 Sem Tratamento (LDI)

Esse método é referido como um tratamento controle, ele manifesta dados que realmente estão presentes nas digitais pois, diferentemente dos demais tratamentos, esse método não apresenta nenhum material exógeno para o auxílio da ionização, sendo assim não há interação das moléculas das digitais com outros materiais e desse modo somente as substâncias presentes na impressão digital coletada serão captadas nas leituras químicas.

De fato, devido a menor facilidade da ionização das moléculas, a quantidade de ionizações desse tratamento foi menor que nos demais tratamentos e nas ionizações captadas a intensidade também foi baixa comparando com o MALDI e o SALDI.

Para exemplificação de quais íons aparecem com maior frequência na amostra, selecionamos os íons das impressões digitais sem a aplicação de micro líquido que se repetiram em cinco ou mais indivíduos.

Tabela 10 – Quantidade de Ionizações para o LDI

Íon	100	106	109	121	149	165	211	309	323	335	352	354
Frequência	5	5	5	5	5	5	5	5	5	5	5	5

Íon	433	281	112	157	191	268	167	307	368	111	151	312
Frequência	5	6	7	7	7	7	8	8	8	9	9	9

Íon	466	536	116	120	178	319	508	103.9	104.1	125	141
Frequência	9	9	10	10	10	10	10	12	12	13	13

Íon	308	293	295	321	115	284	113	494	522	550
Frequência	14	16	17	17	18	18	20	20	20	20

Tendo em vista que na análise descritiva apresentada as frequências relativas dos íons para cada indivíduo no caso sem tratamento, é possível observar que mesmo sem a aplicação de nenhum micro-liquido alguns fatores foram preponderantes, sendo os íons 113, 494, 522 e 550 que foram detectados em todas as impressões digitais. As intensidades desses íons apresentaram as seguintes médias, desvios padrões e coeficientes de variação denotado por C_v:

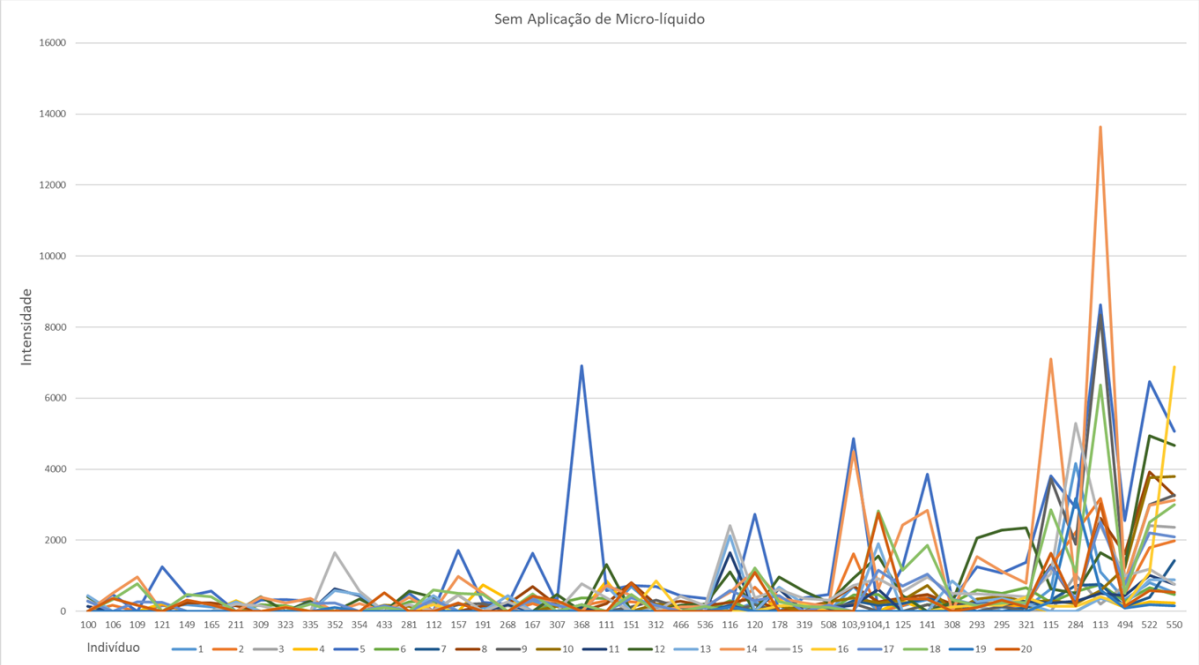
Tabela 11 – Medidas Resumo para o LDI

íon	113	494	522	550
Média	2899.15	680.75	2053.20	2261.80
Desvio Padrão	3603.52	621.34	1709.73	1857.47
C_v	1.243	0.913	0.833	0.821

Os íons que estão presentes somente no sem tratamento e no SALDI são: 211, 323, 112, 157, 167, 307, 368, 111, 151, 536, 116, 120, 319, 508, 125, 141, 308, 293, 321, 115, 113.

Ao analisar o gráfico de perfil dos indivíduos que não tiveram suas impressões digitais submetidas a nenhum tratamento (Figura 9), foi possível constatar que tiveram dois indivíduos com um comportamento próximo e fora dos demais (indivíduos 5 e 14) com picos nos íons (103.9, 113, 115, 125, 141, 157, 293, 295, 321), sendo assim este contraste pode indicar que os mesmos tenham alguma característica em comum. No geral, mesmo que algumas outras impressões tenham um comportamento acentuado em alguns íons, não se teve um padrão de comportamento a ponto de serem consideradas relevantes, as intensidades não são dadas de forma similar ao padrão encontrado no tratamento MALDI.

Figura 9 – Gráfico de perfil para a intensidade dos íons



Algo que se destaca nos três tratamentos para os íons presentes em todos os indivíduos é a alta variabilidade da intensidade desses íons (entre 0.82 e 1.746), a fonte dessa variação pode ser gerado das características dos indivíduos serem diferentes, além do tamanho da amostra ser relativamente pequeno.

4.2 Modelos

Existem dois modos de construção da variável resposta, e com isso tipos diferentes de modelagem. O primeiro modo seria considerar as intensidades dos íons selecionados como as respostas, sendo zero caso o íon não estivesse presente na impressão digital e a intensidade do íon caso esteja presente. A segunda maneira seria considerar como resposta binária, sendo zero a resposta caso o íon não estivesse presente na impressão digital e um caso estivesse presente.

Considerar os dados dessas duas formas diferentes também permite verificar se as variáveis estão somente ligadas a presença ou ausência dos íons ou se a intensidade com que cada íon aparece gera resultados diferentes, além de comparar os modelos em relação a consistência das estimativas dos parâmetros, se independentemente da construção da variável resposta, a significância dos parâmetros continua a mesma.

No primeiro caso, os dados foram modelados via modelos de efeitos mistos, apresentado no Capítulo 2, e no segundo caso os dados foram modelados via modelo misto logístico, apresentado no Capítulo 3.

Para todos os tipos de modelagens testadas, foi comparado o impacto que o tipo de algoritmo de maximização geraria nos modelos. Mesmo sendo apenas vinte pessoas e com muitas medidas dos íons para cada impressão, o resultado obtido foi muito próximo para os algoritmos de Expectation-Maximization, Newton-Raphson e Fisher scoring, com diferença apenas na quinta casa decimal. Optou-se por utilizar os resultados obtidos pelo método de Newton-Raphson, haja vista que é o padrão implementado nos pacotes nlme e lme4 do *Software R*.

Foram analisados os modelos para os três tratamentos, e os dois tipos de abordagens para a variável resposta. Primeiramente, foram construídos modelos com todas as perguntas do questionário, logo após, foram elaborados os modelos com apenas uma covariável para investigar a relação de cada variável explicativa com a resposta. Também foram testados modelos com interações entre covariáveis que se mostraram relevantes nos modelos testados.

Os modelos apresentados consideraram o efeito aleatório somente no intercepto. Para os dados das impressões digitais, seria a mesma correlação entre os íons de um mesmo indivíduo para qualquer massa desses íons, algo viável de supor uma vez que os dados são coletados uma única vez e não há variação visível de correlação entre esses íons. Nesse caso, é considerado que os indivíduos apresentam um mesmo padrão de crescimento ou decrescimento linear.

As variáveis coletadas no questionário são considerados efeitos fixos pois são representadas por um conjunto finito de possíveis respostas, são as próprias respostas avaliadas no experimento, desse modo se tem um efeito médio gerado pelas covariáveis do questionário.

O modelo misto composto por todas as covariáveis do questionário é expresso por:

$$\begin{aligned} y_i = \beta_0 + \beta_1 Idade_i + \beta_2 Gênero_i + \beta_3 Fumante_i + \gamma_1 Anticoncepcional_i + \\ \gamma_2 Antidepressivo_i + \gamma_3 Dipirona_i + \beta_4 Banho_i + \beta_5 XCS_i + \beta_6 CPPDM_i + \\ \beta_7 Limpeza_i + \beta_8 Café_i + b_i + \epsilon_i \end{aligned} \tag{4.1}$$

já o modelo logístico misto composto por todas as covariáveis do questionário é expresso por:

$$\begin{aligned} \ln\left[\frac{p_{ij}}{1-p_{ij}}\right] = \beta_0 + \beta_1 Idade_i + \beta_2 Gênero_i + \beta_3 Fumante_i + \gamma_1 Anticoncepcional_i + \\ \gamma_2 Antidepressivo_i + \gamma_3 Dipirona_i + \beta_4 Banho_i + \beta_5 XCS_i + \beta_6 CPPDM_i + \\ \beta_7 Limpeza_i + \beta_8 Café_i + b_i + \epsilon_{ij} \end{aligned} \tag{4.2}$$

onde os efeitos fixos são as perguntas do questionário: Idade, Gênero, Fumante, Medicamento (subdividido em três grupos, anticoncepcional, antidepressivo e dipirona sódica), Banho, XCS, CPPDM, Limpeza e Café. A resposta sim para essas variáveis foi atribuída como 1 e a resposta não foi atribuída como 0. Para a variável gênero, foi

atribuído 1 para o sexo masculino e 0 para o sexo feminino. O efeito aleatório é denotado por b_i.

Nos modelos estimados que serão apresentados, será dada a estimativa dos parâmetros de efeitos fixos, os seus respectivos erros padrões, seus graus de liberdade assim como o p-valor e a estatística associada ao teste-t para significância destes parâmetros, teste-t esse explicitado na Seção 2.6, onde se testa a hipótese nula de que $\beta_j = 0$. Para os modelos logísticos mistos, será apresentado o p-valor e a estatística associada ao teste de Wald para a significância dos parâmetros, esse teste está explicitado na Seção 3.1.1

Para a estimação desses modelos foi utilizado o programa R 3.5.1. E todas as análises são baseadas em um nível de significância de 10%, portanto para os testes apresentados, tem-se $\alpha = 0.10$. Os modelos com todas as covariáveis, com uma covariável e determinadas interações serão comentados separadamente para cada tratamento nas seções a seguir.

Para todo caso, vale lembrar que os modelos estimados foram baseados em uma amostra de 20 indivíduos e é desejável que se tivesse uma amostra maior. Contudo, a análise feita é um passo inicial dado.

4.2.1 α-matriz (MALDI-TOF)

O modelo misto completo estimado para o tratamento do MALDI é descrito a seguir, onde somente a covariável Limpeza é significativamente diferente de zero pelo teste-t.

Tabela 12 – Modelo misto estimado com todas as covariáveis para o MALDI

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	9314.554	5873.023	1320	1.586	0.113
Idade	301.044	474.886	8	0.634	0.544
Gênero	-8153.740	7201.424	8	-1.132	0.290
Fumante	-3220.639	5748.916	8	-0.560	0.591
Anticoncepcional	-4352.334	4718.224	8	-0.922	0.383
Antidepressivo	-6866.365	5045.239	8	-1.361	0.211
Dipirona Sódica	5405.041	6431.306	8	0.840	0.425
Banho	-6228.146	4307.522	8	-1.446	0.186
XCS	-2016.771	4247.340	8	-0.475	0.648
CPPDM	-7409.627	5555.209	8	-1.334	0.219
Limpeza	12702.893	3534.085	8	3.594	0.007
Café	-3166.463	5830.000	8	-0.543	0.602

Os modelos de efeitos mistos para o tratamento de MALDI-TOF com uma covariável apresentaram em maior parte os efeitos fixos não significativos, onde somente os modelos com banho e café como a covariável explicativa tiveram os efeitos fixos significativos.

Tabela 13 – Modelo misto estimado para o MALDI com uma covariável: Medicamento

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	3295.542	2315.211	1320	1.423	0.155
Anticoncepcional	4559.252	4630.423	16	0.985	0.339
Antidepressivo	-311.378	6125.473	16	-0.051	0.960
Dipirona Sódica	7087.032	6125.473	16	1.157	0.264

Tabela 14 – Modelo misto estimado para o MALDI com uma covariável: Fumante

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	4303.369	1843.861	1320	2.334	0.020
Fumante	5815.892	5830.801	18	0.997	0.332

Tabela 15 – Modelo misto estimado para o MALDI com uma covariável: XCS

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	7851.995	3172.466	1320	2.475	0.013
XCS	-4238.624	3791.822	18	-1.118	0.278

Tabela 16 – Modelo misto estimado para o MALDI com uma covariável: CPPDM

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	8738.162	3095.944	1320	2.822	0.005
CPPDM	-5504.576	3700.361	18	-1.488	0.154

Tabela 17 – Modelo misto estimado para o MALDI com uma covariável: Idade

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	6075.836	3692.380	1320	1.646	0.100
Idade	-130.151	352.938	18	-0.369	0.717

Tabela 18 – Modelo misto estimado para o MALDI com uma covariável: Banho

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	9312.500	2068.769	1320	4.501	<0.001
Banho	-8855.084	2925.681	18	-3.027	0.007

Tabela 19 – Modelo misto estimado para o MALDI com uma covariável: Limpeza

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	437.116	3383.848	1320	0.129	0.897
Limpeza	5930.456	3907.331	18	1.518	0.146

Tabela 20 – Modelo misto estimado para o MALDI com uma covariável: Café

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	9314.224	2068.359	1320	4.503	<0.001
Café	-8858.531	2925.101	18	-3.028	0.007

Tabela 21 – Modelo misto estimado para o MALDI com uma covariável: Gênero

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	5880.272	2519.495	1320	2.334	0.020
Gênero	-1990.627	3563.105	18	-0.559	0.583

A partir desses modelos, pensou-se na possibilidade de que a variável banho poderia retirar o efeito das demais variáveis que representariam substâncias que entram em contato com a pele. Com essa suspeita, foram verificados os modelos considerando a interação do banho com as demais variáveis. Além de considerar a interação da variável banho, também foram considerados as interações da variável XCS com as demais. Considerar as interações da variável XCS foi pensado porque o uso de xampu ou condicionador ou sabonete também poderia retirar o efeito das demais covariáveis.

O único modelo que apresentou a interação significativa foi o modelo com banho, limpeza e a interação entre essas duas covariáveis.

Tabela 22 – Modelo misto final estimado para o tratamento MALDI

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	403.433	1941.468	1320	0.208	0.835
Banho	168.418	4341.255	16	0.039	0.970
Limpeza	14848.445	2506.425	16	5.924	<0.001
Banho:Limpeza	-14975.595	4799.438	16	-3.120	0.007

Esse modelo estimado indica que, utilizar produto de limpeza frequentemente aumenta a resposta média em 14848.445 unidades e a interação do banho com o produto de limpeza diminui a resposta média em 14975.595 unidades. O efeito do banho não é considerado significativo sozinho. Os desvios padrões do resíduo e intercepto são 20152.24 e 3002.635 respectivamente. A correlação entre as medidas repetidas dos indivíduos foi de apenas 0.0217.

O uso frequente de produto de limpeza é bastante relevante para a intensidade dos íons, salvo os casos onde pessoas tomaram banho antes da coleta de suas impressões digitais, o que poderia ser explicado pelo fato de os resquícios desses produtos serem removidos da pele, e assim não estarem mais presentes na impressão digital. As impressões das pessoas que utilizam produto de limpeza frequentemente e que não tomaram banho pela manhã se destacam bastante dos demais indivíduos no tratamento MALDI-TOF pela análise

descritiva, essa diferença se dá ao fato de essas digitais apresentarem uma quantidade maior de ionizações e uma intensidade média de íons muito superior aos demais indivíduos. Essa diferença na ionização é captada pelos modelos e realmente sugerem que esses indivíduos possuam características que os diferem dos demais.

O modelo final com o banho, uso frequente de produto de limpeza e a interação entre eles é proposto para o tratamento MALDI considerando a intensidade. Cabe salientar que embora o café seja significativo quando se tem somente essa variável como covariável explicativa, quando adicionamos essa covariável no modelo com o banho e o produto de limpeza e a interação entre eles, o café deixa de ser significativo, possivelmente pelo efeito dessas variável estar sendo confundida com o efeito produto de limpeza, visto que essas variáveis apresentaram respostas similares na amostra.

No modelo logístico misto, as covariáveis Fumante, Banho, XCS, CPPDM e Limpeza foram significantes, assim como mostra o modelo a seguir.

Tabela 23 – Modelo logístico misto estimado com todas as covariáveis para o MALDI

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	2.076	0.448	4.638	<0.001
Idade	0.006	0.043	0.136	0.892
Gênero	-0.989	0.582	-1.697	0.090
Fumante	-1.715	0.472	-3.633	<0.001
Anticoncepcional	-0.351	0.351	-0.999	0.318
Antidepressivo	-0.182	0.373	-0.487	0.626
Dipirona Sódica	0.339	0.514	0.660	0.509
Banho	-1.784	0.331	-5.389	<0.001
XCS	-1.403	0.337	-4.168	<0.001
CPPDM	-1.634	0.447	-3.655	<0.001
Limpeza	2.263	0.316	7.154	<0.001
Café	0.485	0.428	1.132	0.258

Assim como nos modelos de efeitos mistos, os modelos logísticos mistos com uma covariável explicativa apresentaram em maior parte os efeitos fixos não significativos, mas nessa abordagem, além dos modelos com banho e café, o modelo com Limpeza e XCS como as covariáveis explicativas tiveram os efeitos fixos significativos.

Tabela 24 – Modelo logístico misto estimado para o MALDI com uma covariável: Medicamento

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	-0.049	0.364	-0.134	0.894
Anticoncepcional	1.169	0.736	1.589	0.112
Antidepressivo	0.558	0.957	0.583	0.560
Dipirona Sódica	0.676	0.959	0.705	0.481

Tabela 25 – Modelo logístico misto estimado para o MALDI com uma covariável: Fumante

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	0.325	0.319	1.019	0.308
Fumante	-0.145	1.005	-0.145	0.885

Tabela 26 – Modelo logístico misto estimado para o MALDI com uma covariável: CPPDM

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	0.991	0.533	1.858	0.063
CPPDM	-0.967	0.633	-1.527	0.127

Tabela 27 – Modelo logístico misto estimado para o MALDI com uma covariável: Banho

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	1.083	0.353	3.069	0.002
Banho	-1.548	0.493	-3.142	0.002

Tabela 28 – Modelo logístico misto estimado para o MALDI com uma covariável: Limpeza

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	-0.493	0.561	-0.880	0.379
Limpeza	1.071	0.650	1.648	0.099

Tabela 29 – Modelo logístico misto estimado para o MALDI com uma covariável: Café

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	1.066	0.357	2.983	0.003
Café	-1.512	0.499	-3.031	0.002

Tabela 30 – Modelo logístico misto estimado para o MALDI com uma covariável: Gênero

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	0.543	0.420	1.293	0.196
Gênero	-0.469	0.593	-0.791	0.429

Tabela 31 – Modelo logístico misto estimado para o MALDI com uma covariável: XCS

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	1.105	0.514	2.150	0.032
XCS	-1.137	0.611	-1.859	0.063

Tabela 32 – Modelo logístico misto estimado para o MALDI com uma covariável: Idade

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	0.821	0.611	1.345	0.179
Idade	-0.056	0.058	-0.961	0.337

Assim como a suspeita mencionada nos modelos mistos, para o modelo logístico misto também foram testadas as mesmas interações. Novamente, o único modelo em que a interação foi significativa foi o modelo com banho, limpeza e a interação entre essas duas

covariáveis.

Tabela 33 – Modelo logístico misto final estimado para o tratamento MALDI

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	-0.536	0.243	-2.202	0.028
Banho	0.263	0.540	0.487	0.627
Limpeza	2.673	0.340	7.858	<0.001
Banho:Limpeza	-2.873	0.612	-4.692	<0.001

A interpretação para as estimativas dos modelos mistos e logísticos mistos para o tratamento do MALDI se dá de forma próxima, em que o produto de limpeza geraria, em média, um aumento de 2.673 unidades no logito da variável resposta, enquanto a interação entre o banho e limpeza estaria responsável por diminuir a resposta média do logito em 2.873 unidades, já o banho não foi significativo sozinho. A variância do intercepto para o modelo apresentado na Tabela 33 é de 0.171. Em termos práticos, o produto de limpeza estaria responsável por adicionar íons às impressões digitais, e o banho seria responsável pela retirada desses íons da impressão digital.

Os modelos para o tratamento MALDI mostraram os efeitos fixos significativos para o produto de limpeza e a interação entrem o banho e o produto de limpeza, sendo esses modelos com essas covariáveis considerados adequados.

Esses modelos finais apresentaram AIC de 1530.20 para o modelo logístico misto e 30170.28 para o modelo de efeitos mistos, assim como demonstra a Tabela 34 a seguir.

Tabela 34 – AIC's dos modelos finais para o tratamento MALDI

	Modelo Misto	Modelo Logístico Misto
AIC	30170.28	1530.20

De forma geral, com os dois modelos para o MALDI, é possível diferenciar pessoas que utilizam produto de limpeza frequentemente e que não tomaram banho pela manhã das demais pessoas.

4.2.2 Pó magnético/Sílica (SALDI-MS)

O modelo misto completo para o tratamento SALDI apresentou somente uma covariável significativa, o banho.

Tabela 35 – Modelo misto estimado com todas as covariáveis para o SALDI

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	343.797	447.378	820	0.768	0.442
Idade	-53.168	36.174	8	-1.470	0.180
Gênero	471.408	548.569	8	0.859	0.415
Fumante	49.921	437.924	8	0.114	0.912
Anticoncepcional	-36.991	359.411	8	-0.103	0.921
Antidepressivo	-69.089	384.322	8	-0.180	0.862
Dipirona Sódica	256.512	489.905	8	0.524	0.615
Banho	661.446	328.126	8	2.016	0.079
XCS	336.289	323.541	8	1.039	0.329
CPPDM	296.006	423.169	8	0.699	0.504
Limpeza	75.190	269.209	8	0.279	0.787
Café	-789.856	444.101	8	-1.779	0.113

Dos modelos estimados com uma covariável, nenhum modelo obteve seus efeitos fixos significativos.

Tabela 36 – Modelo misto estimado para o SALDI com uma covariável: Medicamento

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	534.490	102.400	820	5.220	<0.001
Anticoncepcional	-88.073	204.800	16	-0.430	0.673
Antidepressivo	37.903	270.925	16	0.140	0.890
Dipirona Sódica	234.736	270.925	16	0.866	0.399

Tabela 37 – Modelo misto estimado para o SALDI com uma covariável: Fumante

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	525.418	80.362	820	6.538	<0.001
Fumante	187.213	254.126	18	0.737	0.471

Tabela 38 – Modelo misto estimado para o SALDI com uma covariável: XCS

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	638.960	138.725	820	4.606	<0.001
XCS	-135.459	165.808	18	-0.817	0.425

Tabela 39 – Modelo misto estimado para o SALDI com uma covariável: CPPDM

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	579.127	140.929	820	4.109	<0.001
CPPDM	-49.982	168.442	18	-0.297	0.770

Tabela 40 – Modelo misto estimado para o SALDI com uma covariável: Idade

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	653.396	156.860	820	4.165	<0.001
Idade	-11.941	14.994	18	-0.796	0.436

Tabela 41 – Modelo misto estimado para o SALDI com uma covariável: Banho

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	494.505	108.172	820	4.571	<0.001
Banho	99.269	152.978	18	0.649	0.525

Tabela 42 – Modelo misto estimado para o SALDI com uma covariável: Limpeza

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	595.657	154.121	820	3.865	<0.001
Limpeza	-68.690	177.963	18	-0.386	0.704

Tabela 43 – Modelo misto estimado para o SALDI com uma covariável: Café

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	591.645	108.278	820	5.464	<0.001
Café	-95.012	153.128	18	-0.620	0.543

Tabela 44 – Modelo misto estimado para o SALDI com uma covariável: Gênero

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	518.638	109.099	820	4.754	<0.001
Gênero	51.002	154.289	18	0.331	0.745

Do mesmo modo que nos tratamentos anteriores, também foi testada a interação do banho com as demais covariáveis. Os modelos em que o efeito da interação foi significativa foram os que continham o banho e o contato com produto de limpeza com sua interação e o modelo com banho e fumante, assim como sua interação. Cruzando as informações do questionário com a modelagem, optou-se por considerar o modelo da interação do banho e fumante como um modelo não consistente, pois as duas pessoas que fumam não apresentam similaridade no questionário e a suspeita de que o banho retiraria as moléculas do cigarro não foi validada.

O modelo misto final estabelecido para o SALDI considerando a intensidade como variável resposta foi com o banho, contato com produto de limpeza frequentemente e a interação entre essas duas covariáveis. Esse modelo é o mesmo ao estabelecido para o tratamento do MALDI. Suas estimativas são dadas a seguir.

Tabela 45 – Modelo misto final estimado para o tratamento SALDI

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	431.827	158.118	820	2.731	0.006
Banho	819.149	353.563	16	2.317	0.034
Limpeza	104.462	204.130	16	0.512	0.616
Banho:Limpeza	-834.687	390.879	16	-2.135	0.049

A interpretação para este modelo estaria considerando que o banho produz um aumento médio da variável resposta, enquanto a interação entre o banho e limpeza estaria responsável diminuir a resposta média das intensidades, diferentemente do MALDI, que apresentou a limpeza como efeito fixo significativo que seria responsável por aumentar a resposta média da intensidade dos íons. Os desvios padrões do resíduo e intercepto são 1361.755 e 236.3337 respectivamente. A correlação entre as medidas repetidas dos indivíduos foi de apenas 0.0292.

Quando se fez uso do modelo logístico misto, o resultado se mostrou bem semelhante ao modelo misto em que somente um efeito fixo foi significativo, nesse caso a covariável foi Limpeza.

Tabela 46 – Modelo logístico misto estimado com todas as covariáveis para o SALDI

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	0.135	0.419	0.322	0.747
Idade	0.019	0.034	0.550	0.583
Gênero	0.218	0.514	0.425	0.671
Fumante	0.044	0.412	0.106	0.915
Anticoncepcional	0.665	0.338	1.969	0.049
Antidepressivo	0.206	0.363	0.567	0.571
Dipirona Sódica	-0.551	0.460	-1.197	0.231
Banho	-0.443	0.307	-1.445	0.148
XCS	-0.475	0.304	-1.563	0.118
CPPDM	-0.019	0.397	-0.049	0.961
Limpeza	-0.519	0.253	-2.054	0.040
Café	0.670	0.417	1.605	0.109

Assim como no modelo misto, dos modelos logísticos mistos estimados com uma covariável, nenhum modelo obteve seus efeitos fixos significativos.

Tabela 47 – Modelo logístico misto estimado para o SALDI com uma covariável: Medicamento

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	-0.081	0.125	-0.648	0.517
Anticoncepcional	0.226	0.249	0.907	0.364
Antidepressivo	-0.365	0.333	-1.095	0.274
Dipirona Sódica	-0.320	0.334	-0.959	0.338

Tabela 48 – Modelo logístico misto estimado para o SALDI com uma covariável: Fumante

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	-0.060	0.106	-0.562	0.574
Fumante	-0.444	0.341	-1.303	0.193

Tabela 49 – Modelo logístico misto estimado para o SALDI com uma covariável: CPPDM

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	-0.081	0.191	-0.424	0.671
CPPDM	-0.032	0.228	-0.141	0.888

Tabela 50 – Modelo logístico misto estimado para o SALDI com uma covariável: Banho

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	-0.118	0.148	-0.800	0.424
Banho	0.030	0.209	0.144	0.886

Tabela 51 – Modelo logístico misto estimado para o SALDI com uma covariável: Limpeza

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	0.058	0.205	0.284	0.776
Limpeza	-0.215	0.236	-0.911	0.362

Tabela 52 – Modelo logístico misto estimado para o SALDI com uma covariável: Café

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	-0.147	0.147	-1.000	0.317
Café	0.089	0.209	0.425	0.671

Tabela 53 – Modelo logístico misto estimado para o SALDI com uma covariável: Gênero

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	-0.158	0.147	-1.074	0.283
Gênero	0.110	0.208	0.527	0.598

Tabela 54 – Modelo logístico misto estimado para o SALDI com uma covariável: XCS

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	-0.017	0.190	-0.089	0.929
XCS	-0.123	0.227	-0.542	0.588

Tabela 55 – Modelo logístico misto estimado para o SALDI com uma covariável: Idade

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	-0.176	0.216	-0.814	0.416
Idade	0.008	0.021	0.384	0.701

Ao analisar os modelos logísticos mistos com as interações propostas anteriormente, não se obteve-ve efeitos fixos das interações significativas. A fim de replicar o resultado obtido considerando a intensidade dos íons, é representado a seguir o modelo com banho,

limpeza e a interação entre essas covariáveis, mesmo não sendo significativas.

Tabela 56 – Modelo logístico misto final estimado para o tratamento SALDI

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	0.121	0.222	0.544	0.587
Banho	-0.314	0.498	-0.632	0.527
Limpeza	-0.399	0.288	-1.388	0.165
Banho:Limpeza	0.517	0.550	0.939	0.348

Do contrário do que se esperava, os modelos para o tratamento a base do SALDI quase não apresentaram estimativas estatisticamente significativas com os dados das analises químicas, portanto não foi possível distinguir características associadas as perguntas do questionário dos indivíduos da amostra, com base no modelo misto logístico. A variância do intercepto para o modelo apresentado na Tabela 56 é de 0.3539. Os AIC's dos modelos, logístico misto e misto foram 1162.20 e 14477.15 respectivamente, assim como mostrado na Tabela 57.

Tabela 57 – AIC's dos modelos finais para o tratamento SALDI

	Modelo Misto	Modelo Logístico Misto
AIC	14477.15	1162.20

4.2.3 Sem Tratamento (LDI)

No tocante ao sem material, temos o modelo misto estimado com todas as covariáveis, em que quase todos os efeitos fixos foram significativos. Em um primeiro momento esse resultado foi bastante animador, mas ao analisar os modelos com uma covariável explicativa, percebeu-se que possivelmente esses efeitos estão sendo considerados significativos pelo fato de adicionar variáveis no modelo e que esses efeitos não necessariamente são consistentes nos modelos testados, levando em conta que o tamanho da amostra é relativamente pequeno.

Tabela 58 – Modelo misto estimado com todas as covariáveis para o LDI

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	1152.492	264.341	880	4.360	<0.001
Idade	17.764	21.374	8	0.831	0.430
Gênero	-1055.908	324.131	8	-3.258	0.012
Fumante	-739.522	258.755	8	-2.858	0.021
Anticoncepcional	-1142.891	212.364	8	-5.382	0.001
Antidepressivo	-956.105	227.083	8	-4.210	0.003
Dipirona Sódica	348.721	289.469	8	1.205	0.263
Banho	465.482	193.879	8	2.401	0.043
XCS	-32.236	191.170	8	-0.169	0.870
CPPDM	-660.888	250.036	8	-2.643	0.030
Limpeza	517.532	159.067	8	3.254	0.012
Café	-268.561	262.405	8	-1.023	0.336

Dos modelos mistos estimados com uma covariável, nenhum modelo teve seu efeito fixo significativo.

Tabela 59 – Modelo misto estimado para o LDI com uma covariável: Medicamento

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	538.615	98.989	880	5.441	<0.001
Anticoncepcional	-317.398	197.978	16	-1.603	0.128
Antidepressivo	-303.893	261.900	16	-1.160	0.263
Dipirona Sódica	-195.126	261.900	16	-0.745	0.467

Tabela 60 – Modelo misto estimado para o LDI com uma covariável: Fumante

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	422.131	84.004	880	5.025	<0.001
Fumante	31.025	265.643	18	0.117	0.908

Tabela 61 – Modelo misto estimado para o LDI com uma covariável: XCS

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	413.922	145.519	880	2.844	0.005
XCS	16.159	173.928	18	0.093	0.927

Tabela 62 – Modelo misto estimado para o LDI com uma covariável: CPPDM

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	567.444	139.931	880	4.055	<0.001
CPPDM	-203.159	167.250	18	-1.215	0.240

Tabela 63 – Modelo misto estimado para o LDI com uma covariável: Idade

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	442.203	164.371	880	2.690	0.007
Idade	-1.855	15.711	18	-0.118	0.907

Tabela 64 – Modelo misto estimado para o LDI com uma covariável: Banho

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	411.633	112.654	880	3.654	<0.001
Banho	27.200	159.317	18	0.171	0.866

Tabela 65 – Modelo misto estimado para o LDI com uma covariável: Limpeza

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	357.867	158.388	880	2.259	0.024
Limpeza	89.822	182.891	18	0.491	0.629

Tabela 66 – Modelo misto estimado para o LDI com uma covariável: Café

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	459.418	112.168	880	4.096	<0.001
Café	-68.369	158.630	18	-0.431	0.672

Tabela 67 – Modelo logístico misto estimado para o LDI com uma covariável: Gênero

	Estimativa	Erro Padrão	g.l.	Valor t	p-valor
Intercepto	408.087	112.600	880	3.624	<0.001
Gênero	34.293	159.241	18	0.215	0.832

Do mesmo modo que foi testado no tratamento MALDI, as interações do banho e as interações do XCS para o LDI também foram testadas. Desta forma notou-se que assim como os modelos com uma covariável indicavam, nenhuma interação foi significativa e portanto não foi proposto um modelo final considerando a intensidade.

Os modelos logísticos mistos estimados com uma covariável e completos estão apresentados a seguir.

Tabela 68 – Modelo logístico misto estimado com todas as covariáveis para o LDI

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	1.185	0.596	1.989	0.047
Idade	0.018	0.048	0.374	0.709
Gênero	-1.465	0.741	-1.977	0.048
Fumante	-0.703	0.581	-1.211	0.226
Anticoncepcional	-1.723	0.488	-3.532	<0.001
Antidepressivo	-1.567	0.514	-3.050	0.002
Dipirona Sódica	0.478	0.648	0.739	0.460
Banho	0.584	0.433	1.351	0.177
XCS	-0.123	0.426	-0.289	0.773
CPPDM	-1.086	0.562	-1.932	0.053
Limpeza	0.725	0.356	2.038	0.042
Café	-0.213	0.582	-0.367	0.714

Tabela 69 – Modelo logístico misto estimado para o LDI com uma covariável: Medicamento

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	0.186	0.186	1.004	0.315
Anticoncepcional	-0.605	0.371	-1.632	0.103
Antidepressivo	-0.752	0.492	-1.529	0.126
Dipirona Sódica	-0.186	0.492	-0.379	0.704

Tabela 70 – Modelo logístico misto estimado para o LDI com uma covariável: Fumante

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	-0.080	0.164	-0.488	0.626
Fumante	0.502	0.516	0.972	0.331

Tabela 71 – Modelo logístico misto estimado para o LDI com uma covariável: CPPDM

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	0.298	0.275	1.084	0.279
CPPDM	-0.469	0.330	-1.424	0.154

Tabela 72 – Modelo logístico misto estimado para o LDI com uma covariável: Banho

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	-0.045	0.224	-0.200	0.841
Banho	0.031	0.318	0.098	0.922

Tabela 73 – Modelo logístico misto estimado para o LDI com uma covariável: Limpeza

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	-0.102	0.316	-0.324	0.746
Limpeza	0.098	0.365	0.267	0.789

Tabela 74 – Modelo logístico misto estimado para o LDI com uma covariável: Café

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	0.021	0.224	0.093	0.926
Café	-0.100	0.317	-0.317	0.751

Tabela 75 – Modelo logístico misto estimado para o LDI com uma covariável: Gênero

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	-0.117	0.222	-0.524	0.600
Gênero	0.175	0.315	0.555	0.579

Tabela 76 – Modelo logístico misto estimado para o LDI com uma covariável: XCS

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	0.031	0.289	0.107	0.915
XCS	-0.086	0.346	-0.250	0.803

Tabela 77 – Modelo logístico misto estimado para o LDI com uma covariável: Idade

	Estimativa	Erro Padrão	Valor z	p-valor
Intercepto	0.028	0.327	0.087	0.931
Idade	-0.006	0.031	-0.202	0.840

As conclusões para os modelos logísticos mistos se dão de forma muito próxima ao observado nos modelos mistos. Até mesmo as variáveis banho e XCS com suas interações não apresentaram efeitos fixos significativos.

Para esse tratamento, não foi possível associar os valores da análise química com as respostas dos questionário. Os modelos testados não foram conclusivos e não é viável propor um modelo logístico misto ou um modelo misto para os dados.

4.3 Análise das Suposições do Modelo

Como visto no Capítulo 2, os pressupostos para o modelo de efeitos mistos são: normalidade dos resíduos e do efeito aleatório, independência entre as respostas dos indivíduos da amostra.

Para testar a normalidade do efeito aleatório e dos resíduos, o teste de Shaphiro-Wilk para normalidade foi realizado. Para todos os modelos de efeitos mistos finais, os p-valores dos testes de Shapiro-Wilk, tanto para os resíduos e tanto para os efeitos aleatórios, foram muito próximos de zero, indicando que nem o efeito aleatório nem os resíduos possuem distribuição normal.

O desvio do pressuposto de normalidade é visível pelo fato de o p-valor ser muito próximo de zero no modelos mistos para o MALDI e o SALDI. Análises gráficas também indicaram que esse pressuposto não é satisfeito. Possivelmente a falta de normalidade advém da alta variabilidade dos dados e da inflação das respostas no zero. Assim sendo, os modelos mistos não são ideais para esses dados em relação a suposição de normalidade dos dados. Esse é um ponto motivador para que os modelos logísticos mistos fossem propostos.

Como visto no Capítulo 3, os pressupostos para o modelo logístico misto são: apenas a normalidade do efeito aleatório e também a independência entre as respostas dos indivíduos da amostra, assim como no modelo de efeitos mistos.

Já nos modelos logísticos mistos finais estimados, houve uma melhora significativa em relação ao pressuposto de normalidade do efeito aleatório, nos modelos finais, os p-valores para o teste de Shaphiro-Wilk foram 0.1614 e 0.1212 para o SALDI e MALDI respectivamente. Com esses valores para o testes de normalidade, é admissível aceitar que o efeito aleatório dos três modelos logísticos seguem uma distribuição de probabilidade normal, e assim o pressuposto de normalidade é satisfeito.

Em relação a independência entre as respostas dos indivíduos, tomou-se cuidado para que não houvesse contaminação de uma impressão digital com as outras, além do não contato dos indivíduos no dia da coleta. Pode-se afirmar que a suposição de independência entre os indivíduos é razoável dada a natureza do experimento, ao qual as informações obtidas são individuais, e não há uma relação de dependência entre os indivíduos.

O gráfico quantil-quantil (q-q) é um instrumento capaz de auxiliar na comparação da distribuição empírica da variável com a distribuição teórica proposta, a distribuição normal. Serão apresentados somente os gráficos para os efeitos aleatórios dos modelos logísticos, pois como já comentado, os resíduos e o efeitos aleatórios do modelo misto claramente não possuem distribuição normal, portanto os gráficos apresentados nas Figuras 10 e 11 são:

Figura 10 – Gráfico q-q para os Efeitos Aleatórios do SALDI

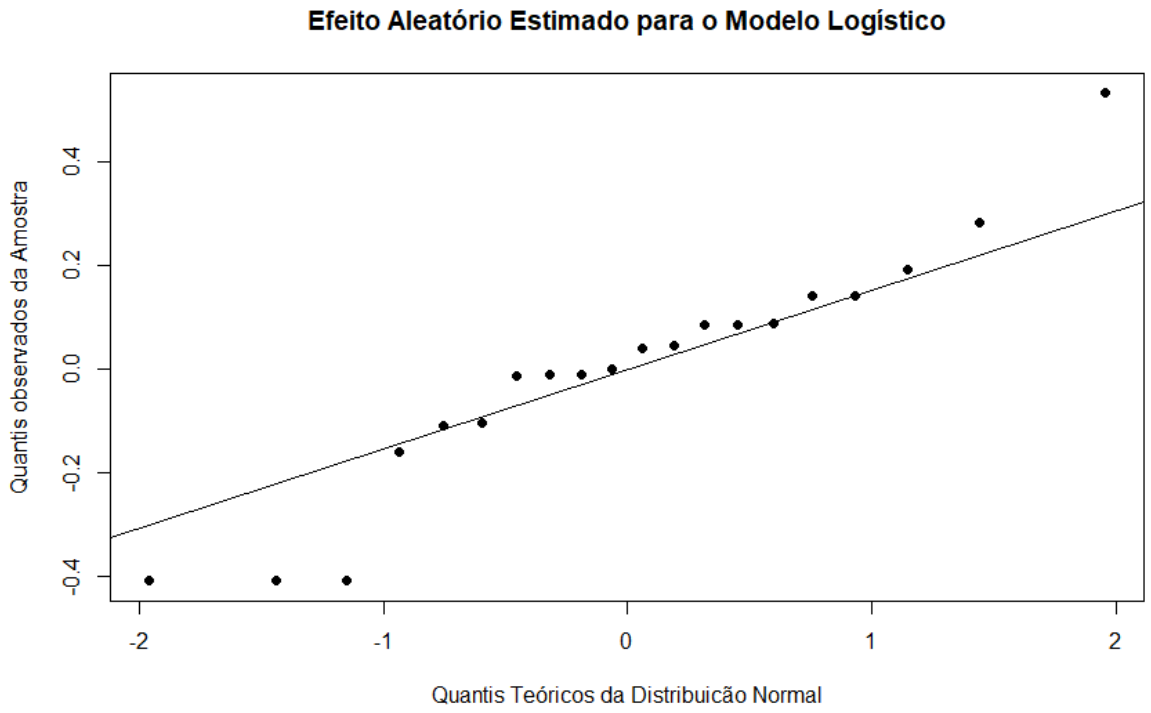
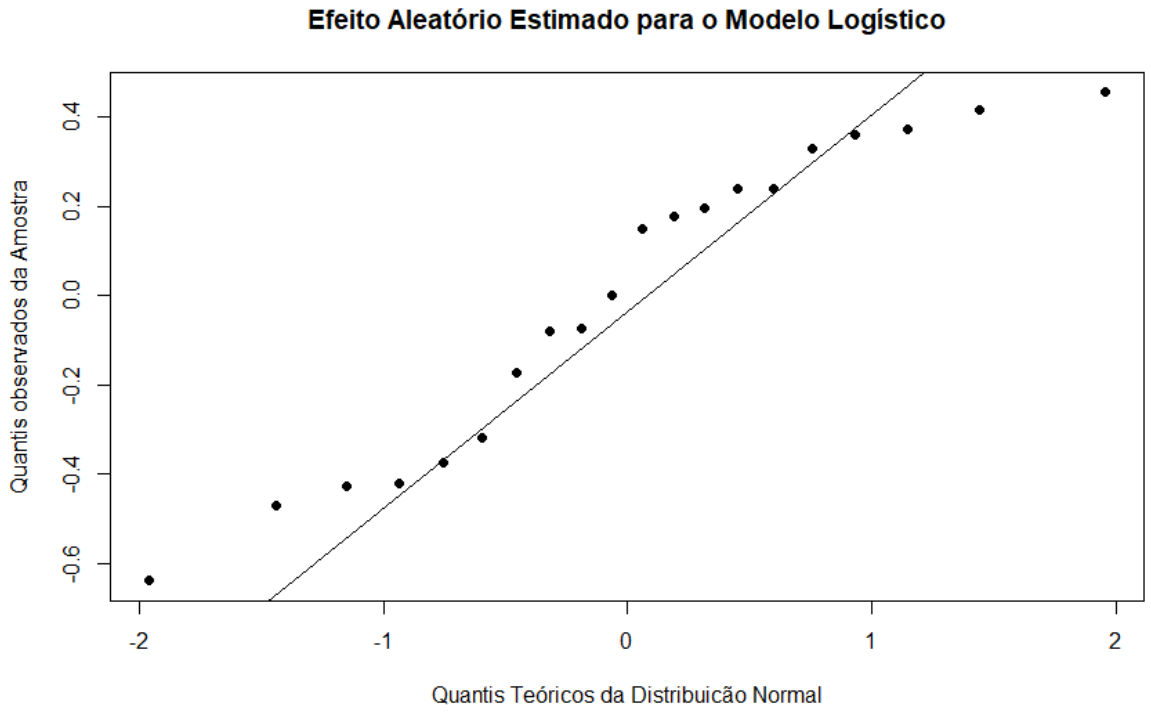


Figura 11 – Gráfico q-q para os Efeitos Aleatórios do MALDI



Nos gráficos com os tratamentos SALDI e MALDI, para o modelo logístico misto, os pontos se apresentaram de forma mais dispersa em torno da reta teórica da distribuição normal, mas ainda assim é possível supor que os efeitos possuem aproximadamente uma distribuição normal, concordando também com o teste de Shaphiro-Wilk.

Conforme a melhor adequabilidade dos pressupostos para a construção dos modelos logísticos mistos, estes se tornam melhores em relação ao modelo misto, ou seja, a implementação desses modelos gerou melhores resultados a serem analisados.

Tabela 78 – Critério AIC para os modelos finais

	MALDI	MALDI log	SALDI	SALDI log
AIC	30170.28	1530.20	14477.15	1162.20

Conforme a Tabela 78, os modelos logísticos apresentaram AIC's consideravelmente inferiores, isso é decorrido por conta da modelagem considerando a intensidade dos íons, onde os pressupostos para o modelo não são satisfeitos. Esse critério indica que a modelagem logística é consideravelmente melhor em relação a resposta, dito isso, possivelmente a presença ou ausência de determinado íon já seria suficiente para identificar as substâncias.

5 Conclusão

Como visto anteriormente, o principal objetivo deste trabalho é identificar por meio de um método estatístico possíveis características dos indivíduos relacionadas aos íons obtidos na analise química das impressões digitais. A partir disso, supondo que os modelos sejam apropriados para a disposição dos dados, pretende-se identificar características de indivíduos a partir da análise de impressões digitais, de forma que seja possível subsidiar estatisticamente estudos nessa área.

Com o desafio de trazer uma nova tecnologia para o Brasil que poderia servir na identificação de cadáveres, suspeitos, crimes, drogas, venenos, explosivos e assim por diante, o novo nano-material não se mostrou eficiente como era o esperado para o experimento. Com os dados que trabalhamos, não ficou estatisticamente comprovado sua eficiência quanto a denuncia almejada dos efeitos nos modelos.

O uso frequente de produto de limpeza se mostrou significante na maioria dos modelos testados, porém somente foi possível associar os modelos com as características do questionário no tratamento MALDI.

O MALDI-TOF, embora que com a suspeita de que a excitação das moléculas seria demasiada e geraria muito ruído na análise química, foi o tratamento que teve uma resposta melhor na modelagem. Ele foi capaz de distinguir pessoas que utilizam produto de limpeza frequentemente e que não tomaram banho pela manhã das demais pessoas, uma vez que as respostas das pessoas que usam produto de limpeza frequentemente e que não tomaram banho são, em média, muito maiores das demais pessoas que não apresentaram essa combinação de característica. Esse resultado é evidenciado tanto na analise descritiva quanto nos modelos testados.

Novas metodologias podem ser testadas para tratar os dados, como por exemplo considerar as variáveis do questionário como as variáveis respostas dos modelos, e assim tentar produzir modelos preditivos que consigam prever as respostas dos questionários.

Esse trabalho é um passo inicial para estudos nessas áreas, visto que não se tem uma vasta literatura estatística sobre essa área, impossibilitando a replicação de algum estudo semelhante apesar de já ser uma técnica conhecida, entretanto para um problema novo. O experimento apresentou algumas dificuldades como a formulação de algumas perguntas do questionário e o tamanho da amostra.

Referências

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Brady, J. J., Judge, E. J., and Levis, R. J. (2009). Mass spectrometry of intact neutral macromolecules using intense non-resonant femtosecond laser vaporization with electrospray post-ionization. *Rapid communications in mass spectrometry*, 23(19):3151–3157.

Casella, G. and Berger, R. L. (2010). Inferência estatística. *São Paulo: Cengage Learning*.

Codeço, A. G. (1991). *Elementos básicos da perícia criminal*. Rio de Janeiro: Lélu.

Correia, L. T. (2007). Análise de resposta muscular por eletromiografia utilizando medidas repetidas/dados longitudinais.

de FREITAS, A. and FILHO, J. S. d. S. B. (2005). Modelos lineares mistos para classificação em jogos de duplas. *Rev. Mat. Estat*, 23(3):19–31.

Demidenko, E. (2013). *Mixed models: theory and applications with R*. John Wiley & Sons.

Duarte, A. L. A. and Carvalho, M. F. S. (2018). Análise de agrupamento em impressões digitais.

Kuehl, R. O. (2001). *Diseño de experimentos: principios estadísticos para el diseño y análisis de investigaciones*. Thomson Learning,.

Nemes, P. and Vertes, A. (2007). Laser ablation electrospray ionization for atmospheric pressure, in vivo, and imaging mass spectrometry. *Analytical chemistry*, 79(21):8098–8106.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554.

Piantino, A. C. F. and Marques, F. d. S. (2014). Modelos lineares mistos aplicados aos dados do programa doce desafio.

Tanaka, K. (2003). The origin of macromolecule ionization by laser irradiation (nobel lecture). *Angewandte Chemie International Edition*, 42(33):3860–3870.

Vilhena, G. J. (2013). Modelos lineares mistos: aplicado na modelagem de níveis de glicemia.

Wieser, A., Schneider, L., Jung, J., and Schubert, S. (2012). Maldi-tof ms in microbiological diagnostics—identification of microorganisms and beyond (mini review). *Applied microbiology and biotechnology*, 93(3):965–974.

Wu, L. (2009). *Mixed effects models for complex data*. CRC Press.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., and Smith, G. M. (2009). Zero-truncated and zero-inflated models for count data. In *Mixed effects models and extensions in ecology with R*, pages 261–293. Springer.

Anexo 1

1. Qual a sua idade?
2. Possui dieta específica? (ex: vegetariano, vegano)
3. É fumante? Se sim, quantos cigarros por dia?
4. Faz uso frequente de algum medicamento? Se sim, qual ou quais?
5. Utilizou algum medicamento nas últimas 24 horas? Se sim, qual ou quais?
6. Tomou banho pela manhã (antes de doar suas impressões)?
7. Utilizou cosméticos como xampus, sabonetes, cremes, protetor solar e/ou maquiagem nas horas que antecederam o exame? Se sim, quais?
8. Possui o hábito de lavar a louça e/ou realizar a faxina em casa sem o uso de luvas?
9. Ingeriu café (bebida) pela manhã (antes de doar suas impressões)?

Anexo 2

```
#Pacotes nlme e lme4
#Modelo misto completo
M <- lme(mz_maldi ~ Idade + Gênero + Fumante + Medicamento
                    + Banho + xcs + cppdm + Limpeza + Café,
                    random = ~1 | Doador, data = df)
#Modelos mistos com uma covariável
M1 <- lme(mz_maldi ~ Medicamento, random = ~1 | Doador, data = df)
M2 <- lme(mz_maldi ~ Fumante, random = ~1 | Doador, data = df)
M3 <- lme(mz_maldi ~ xcs, random = ~1 | Doador, data = df)
M4 <- lme(mz_maldi ~ cppdm, random = ~1 | Doador, data = df)
M5 <- lme(mz_maldi ~ as.numeric(Idade), random = ~1 | Doador, data = df)
M6 <- lme(mz_maldi ~ Banho, random = ~1 | Doador, data = df)
M7 <- lme(mz_maldi ~ Limpeza, random = ~1 | Doador, data = df)
M8 <- lme(mz_maldi ~ Café, random = ~1 | Doador, data = df)
M9 <- lme(mz_maldi ~ Gênero, random = ~1 | Doador, data = df)
#Modelo logístico misto completo
m <- glmer(teste~ Idade + Gênero + Fumante + Medicamento + Banho
                  + xcs + cppdm + Limpeza + Café +(1|Doador),
                  data=dft, family=binomial)
```

```
#Modelos logísticos mistos com uma covariável
m1 <- glmer(teste~Medicamento+(1|Doador), data=dft, family=binomial)
m2 <- glmer(teste~Fumante+(1|Doador), data=dft, family=binomial)
m3 <- glmer(teste~cppdm+(1|Doador), data=dft, family=binomial)
m4 <- glmer(teste~Banho+(1|Doador), data=dft, family=binomial)
m5 <- glmer(teste~Limpeza+(1|Doador), data=dft, family=binomial)
m6 <- glmer(teste~Café+(1|Doador),data=dft, family=binomial)
m7 <- glmer(teste~Gênero+(1|Doador),data=dft, family=binomial)
m8 <- glmer(teste~xcs+(1|Doador),data=dft, family=binomial)
m9 <- glmer(teste~Idade+(1|Doador),data=dft, family=binomial)
```